A Framework for the Development and Interpretation of Different Sepsis Definitions and Clinical Criteria

Derek C. Angus, MD[1]; Christopher W. Seymour, MD[1]; Craig M. Coopersmith, MD[2]; Clifford S. Deutschman, MD[3]; Michael Klompas, MD[4,5]; Mitchell M. Levy, MD[6]; Gregory S. Martin, MD[7]; Tiffany M. Osborn, MD[8]; Chanu Rhee, MD[4,5]; R. Scott Watson, MD[9]

[1]Department of Critical Care Medicine, The Clinical Research, Investigation, and Systems Modeling of Acute illness (CRISMA) Center, University of Pittsburgh School of Medicine, Pittsburgh, PA.

[2]Department of Surgery, Emory University School of Medicine, Atlanta, GA.

[3]Department of Pediatrics, Hofstra-North Shore-LIJ School of Medicine, Cohen Children's Medical Center, New Hyde Park, NY.

[4]Department of Population Medicine, Harvard Pilgrim Health Care Institute and Harvard Medical School, Boston, MA.

[5]Department of Medicine, Brigham and Women's Hospital, Boston, MA.

[6]Division of Pulmonary/Critical Care Medicine, Alpert Medical School at Brown University, Providence, RI.

[7]Division of Pulmonary, Allergy, and Critical Care Medicine, Department of Critical Care, Emory University School of Medicine, Atlanta, GA.

[8]Departments of Surgery and Emergency Medicine, Washington University School of Medicine, St. Louis, MO.

[9]Department of Pediatrics, Pediatric Critical Care Medicine, University of Washington; Center for Child Health Behavior and Development, Seattle Children's Research Institute, Seattle, WA.

Drs. Angus and Seymour and were supported, in part, by grants from the National Institutes of Health (NIH) (GM104022, GM107650, and HL123020). Dr. Seymour received support for article research from the NIH and he received funding from Beckman Coulter. His institution received funding from the NIH. Dr. Coopersmith disclosed other support (He was the president of the Society of Critical Care Medicine [SCCM] when the manuscript was submitted. Salary support for effort was paid to Emory University for this position. He also receives grant support from the NIH paid to Emory University unrelated to this paper. He also receives salary support from the Centers for Disease Control and Prevention [CDC] related to sepsis surveillance). Dr. Deutschman disclosed receiving other support from WFSICC (hotel expenses, meeting PA), SCCM (Honorarium, travel expenses for presentation), and SCCM (reimbursement for attending meetings, accommodations at annual meeting). He received funding from the CDC, North Shore LIJ Health System, and the Department of Anesthesiology, Stanford University. Dr. Klompas' institution received funding from the CDC. Dr. Levy's institution received grant support from ImmuneExpress (Gene Analysis for early identification of sepsis). Dr. Martin received support for article research from the NIH and received funding from CR Bard and Medscape. His institution received funding from the CDC, NIH, FDA, and Baxter Healthcare. Dr. Osborn disclosed receiving other support from the Barnes Jewish Hospital. Her institution received funding from Cheetah Inc. and ImaCor Inc. Dr. Watson disclosed receiving other support from the University of Washington and University of Pittsburgh UPMC and he received funding from the CDC (paid travel funds for a meeting to discuss sepsis monitoring). His institution received funding from the NIH/NICHD, SCCM, and the Seattle Children's Hospital. The remaining authors have disclosed that they do have no potential conflicts of interest.

For information regarding this article, E-mail: angusdc@upmc.edu

DOI: 10.1097/CCM.0000000000001730

Abstract: Although sepsis was described more than 2,000 years ago, and clinicians still struggle to define it, there is no "gold standard," and multiple competing approaches and terms exist. Challenges include the ever-changing knowledge base that informs our understanding of sepsis, competing views on which aspects of any potential definition are most important, and the tendency of most potential criteria to be distributed in at-risk populations in such a way as to hinder separation into discrete sets of patients. We propose that the development and evaluation of any definition or diagnostic criteria should follow four steps: 1) define the epistemologic underpinning, 2) agree on all relevant terms used to frame the exercise, 3) state the intended purpose for any proposed set of criteria, and 4) adopt a scientific approach to inform on their usefulness with regard to the intended purpose. Usefulness can be measured across six domains: 1) reliability (stability of criteria during retesting, between raters, over time, and across settings), 2) content validity (similar to face validity), 3) construct validity (whether criteria measure what they purport to measure), 4) criterion validity (how new criteria fare compared to standards), 5) measurement burden (cost, safety, and complexity), and 6) timeliness (whether criteria are available concurrent with care decisions). The relative importance of these domains of usefulness depends on the intended purpose, of which there are four broad categories: 1) clinical care, 2) research, 3) surveillance, and 4) quality improvement and audit. This proposed methodologic framework is intended to aid understanding of the strengths and weaknesses of different approaches, provide a mechanism for explaining differences in epidemiologic estimates generated by different approaches, and guide the development of future definitions and diagnostic criteria. (*Crit Care Med* 2016; 44:e113–e121)

Key Words: definitions, diagnosis, diagnostic criteria, measurement, reliability, sepsis, validity

Although sepsis was first described more than 2,000 years ago, clinicians and researchers still struggle to define it (1). While the identification of florid meningococcal sepsis may present little difficulty, such classic cases are unusual, and even in these, the diagnosis may not be obvious

until the disease has progressed beyond the optimal time for intervention. With advances in our understanding of the pathophysiology of sepsis and heightened awareness of its public health importance, there is increasing pressure to have widely deployable, consistent, and accurate diagnostic criteria, which in turn sparks a desire for a so-called "valid" definition. However, there are competing definitions and criteria by which sepsis is measured. These different approaches identify different patients and produce different estimates of incidence and outcome, generating frustration and confusion.

The problem is that defining sepsis, like any disease or syndrome, is more complex than might readily be apparent. But complexity necessitates neither opacity nor futility. With transparency and rigor, approaches can be developed to define and measure sepsis. And, there can be room for more than one approach without one being more right than another. Indeed, different approaches can serve different purposes; the key is to understand the relationship between the approaches, such that potential differences in case identification can be predicted and understood. This article is the first of two designed to provide a road map for better understanding of the motivations, strengths, and weaknesses of different sepsis definitions and criteria. In this article, we review the inherent challenges facing sepsis, highlighting that many of these challenges are common to the definition of any disease or syndrome. We then describe ground rules for disease and syndrome classification exercises, emphasizing that it is an elusive task to generate a single all-encompassing definition. In the second article, we apply these concepts to different sepsis definitions and criteria.

WHY DEFINING SEPSIS SEEMS DIFFICULT

The 2016 Third International Consensus Definitions for Sepsis and Septic Shock (2) is "infection complicated by life-threatening organ dysfunction due to a dysregulated host response." Although this definition nicely encapsulates the current thinking about sepsis, it also illustrates key challenges. Let us start by rewriting this definition as a logic statement:

$$sepsis = f\left(threat\ to\ life \mid organ\ dysfunction \mid dysregulated\ host\ response \mid infection\right)$$

where sepsis is a function of four variables linked in a causal pathway with, from left to right, one conditional upon the other. We can probe this statement both regarding its internal structure and its external validity or usefulness.

Internal Structure of the Definition

To probe the structure, we ask whether each variable exists, whether it can be measured, and whether the conditional relationships hold. For example, most might agree that organ dysfunction exists, in that organs exist and they appear to function differently compared to their normal healthy state. However, their functions can be multiple, some of which are still unknown. There is controversy regarding whether deviation from normal behavior during a challenge is always dysfunctional, which implies a pathologic or disadvantageous response, or sometimes an appropriate functional stress response to facilitate recovery. There may also be uncertainty over how to measure organ dysfunction, and even whether it is possible. This uncertainty can be considered in terms of each organ, how one integrates across multiple organs, and whether organs are best framed as traditional anatomic units (such as lung or kidney) or as common functional failures across anatomic organs (such as endothelial leak or mitochondrial dysfunction). Assuming organ dysfunction can be measured, attributing the marginal degradation in function to a dysregulated host response is not trivial, and requires an ability to determine preexisting dysfunction, any noninfectious contributions to dysfunction, and, ideally, the mechanism by which the host response to an infection causes organ dysfunction. Similar questions relate to the other variables and purported causal relationships. For example, what exactly is the host response and can we measure the extent to which it is dysregulated? And, can we trace the mechanism by which the organ dysfunction is due solely to this host response?

External Validity and Usefulness of the Definition

Assuming the logic of the statement holds, we can also ask whether it is useful for advancing science or treating patients and, if not, whether it can be modified or replaced. Most broadly, for example, one can ask: Do we even need a definition for sepsis? Or, is it better to define individual infectious diseases, perhaps grading them by their severity, or to define subsets of patients based on the presence of specific constellations of molecular abnormalities responsible for different portions of the dysregulated host response or organ dysfunction patterns? Alternatively, should the logic statement above be altered? For example, by changing infection to a broader set of acute insults, thus allowing the possibility that sepsis can arise from sterile acute pancreatitis or major trauma.

MANAGING DESPAIR AND REMEMBERING THE GRASS IS RARELY GREENER

Although the above questions seem daunting, two points are worth noting. First, they need not all be answered. We enunciate them to point out that a perfect definition likely requires a perfect understanding, and neither is close at hand. But we need not be perfect, just good. As with all of science, we are simply trying to generate a good working model, or hypothesis: one that integrates current understanding and that can be updated as knowledge accrues. Second, experts often believe that their disease or condition is the hardest to tackle. For example, psychiatrists argue that their disorders cannot be parsed as easily as asthma from emphysema (3). Yet, pulmonologists quickly list the considerable challenges of parsing out various subsets of overlapping obstructive lung diseases (4, 5). Similarly, both groups may consider oncologists to have an easier job diagnosing many cancers, and yet tumor classification undergoes almost constant revision, often with considerable controversy (6). The reason, of course, is that almost all diseases and syndromes are challenging to define in one way or another. With the possible exception of Mendelian-inherited diseases, perfectly discrete and unambiguous disease definitions will likely remain elusive in all areas of medicine for some time to come. Sepsis may be difficult, but many of its problems are not unique.

GOALS AND CHALLENGES FOR ALL DISEASE CLASSIFICATIONS

Defining any disease or syndrome is an exercise in classification (also known as categorization or disambiguation), the process by which ideas and objects are recognized, differentiated, and understood (7). The goal is clear: the convenience that comes with assigning a discrete label: "This patient has disease x." Once labeled, we can count cases, assess effectiveness of different treatments, and measure outcomes. However, three challenges are inherent to virtually all disease classification exercises: overcoming problems of knowledge, purpose, and statistics.

The Knowledge Problem

The knowledge on which physicians and biomedical scientists rely to form theories and opinions about health and disease is constantly changing, incomplete, and variable, resulting in a profusion of theories that may satisfy conditions local to a particular disease, moment in time, or discipline, yet are neither unified nor consistent. For centuries, sepsis was explained primarily by the germ theory as articulated first by Fracastoro and informed by the work of Semmelweis, Pasteur, Lister and others (1, 8). In the late 20th century, with the advent of intensive care and antibiotics, it became apparent that patients with sepsis could die despite eradication of the invading organism (9, 10). This observation led to the host theory, first articulated as systemic inflammatory response syndrome (SIRS) (11). That theory has subsequently been modified with appreciation of complex host-pathogen interactions and variation in the host response. Today, there are proponents to varying degrees for all of these theories.

The Purpose Problem

One's values and priorities shape in important ways how one judges the performance of a particular classification scheme. Broadly speaking, a disease or syndrome classification scheme helps four purposes: clinical care, basic and clinical research, epidemiology and surveillance, and quality improvement and audit. Even with perfect access to information, practitioners of each of these different applications may favor different classification schemes. For example, an immunologist would likely give greater weight to a scheme that divided individuals based on host immune response patterns. Such a scheme would also be useful to clinicians if therapies were based on select immune responses. But because current treatments are initiated largely in response to nonspecific clinical features, clinical diagnostic criteria are likely rated more important by the clinician. Furthermore, the clinician seeks a disease classification that can be applied prospectively to guide treatment decisions. In contrast, an epidemiologist may favor a scheme that most accurately parses cases from non-cases, even if that scheme included postmortem findings.

The Statistical Problem

Disease classification requires separation into discrete sets: for this patient to have disease x, we must say that she is no longer in the set of individuals who do not have disease x, and may further argue that she has disease x as opposed to disease y. Thus, one key property when sorting through a population of individuals is that each can be neatly assigned to one set or another, and few would be in a grey zone of "maybe having disease x." In 1957, Sneath, a bacterial taxonomist, coined this desired property of a classification scheme the "point of rarity," citing that the platypus existed at the point of rarity between birds, mammals, and reptiles. If the platypus were one of many analogous species, the division between birds, mammals and reptiles would not be at a point of rarity and, consequently, feel less useful (12). Kendell and Jablensky, reflecting on disease and syndrome definitions in psychiatry, rephrased the term as the "zone of rarity," and pointed out that many proposed diseases and syndromes are not bound by zones of rarity (3). Rather, the set of clinical and biologic characteristics (so called "surface phenomena" (3) used to define a particular condition is often expressed on a continuum, and the frequency distribution of individuals across the range of these surface phenomena does not usually contain discrete peaks separated by zones of rarity (**Fig. 1, *A–B***).

This problem permeates most criteria one might use to define sepsis. For example, there is no zone of rarity in the population distributions of fever or white blood cell counts that would facilitate discrimination between those with and without infection. Similarly, no zones of rarity exist for common measures of organ dysfunction (**Fig. 1, *C–D***). One exception is the use of discrete interventions, such as intubation, to define organ dysfunction: patients are either intubated or they are not. However, there are other issues with such an approach, as discussed later.

SETTING GROUND RULES FOR DISEASE AND SYNDROME CLASSIFICATION EXERCISES

Beyond the inherent problems discussed above, the methods for classifying diseases and syndromes themselves are also complex and often executed inexpertly (**Table 1**). We can organize classification exercises into four broad components: 1) the epistemologic or philosophical underpinning (e.g., adopting a nominalist versus realist framework regarding the extent to which abstract elements and sets of elements can exist and the criteria for attributes that govern whether an individual element can be grouped or not), 2) agreement over terms and definitions that are not under study but are critical to the framing of the exercise (e.g., if the goal is to define a particular disease, then the definition for the term "disease" must be agreed on at the outset), 3) a prioritization of values that would permit one classification scheme to be judged superior to another (e.g., representativeness of a biologic mechanism may be considered more or less important than ease of timely diagnosis), and 4) the scientific methods by which the exercise will be conducted (e.g., specification of the rationale and approach for all deliberative processes and empiric data analyses).

The stringency with which physicians have delineated these characteristics as they embark on disease or syndrome classification has often been lacking. More than 50 years ago, Eden stated, "The major problems in the methodology of medical diagnosis are problems for the physicians, and until the physician is willing to investigate his [sic] own terminology [and] methodology, all the computer engineer, physical scientist or mathematician

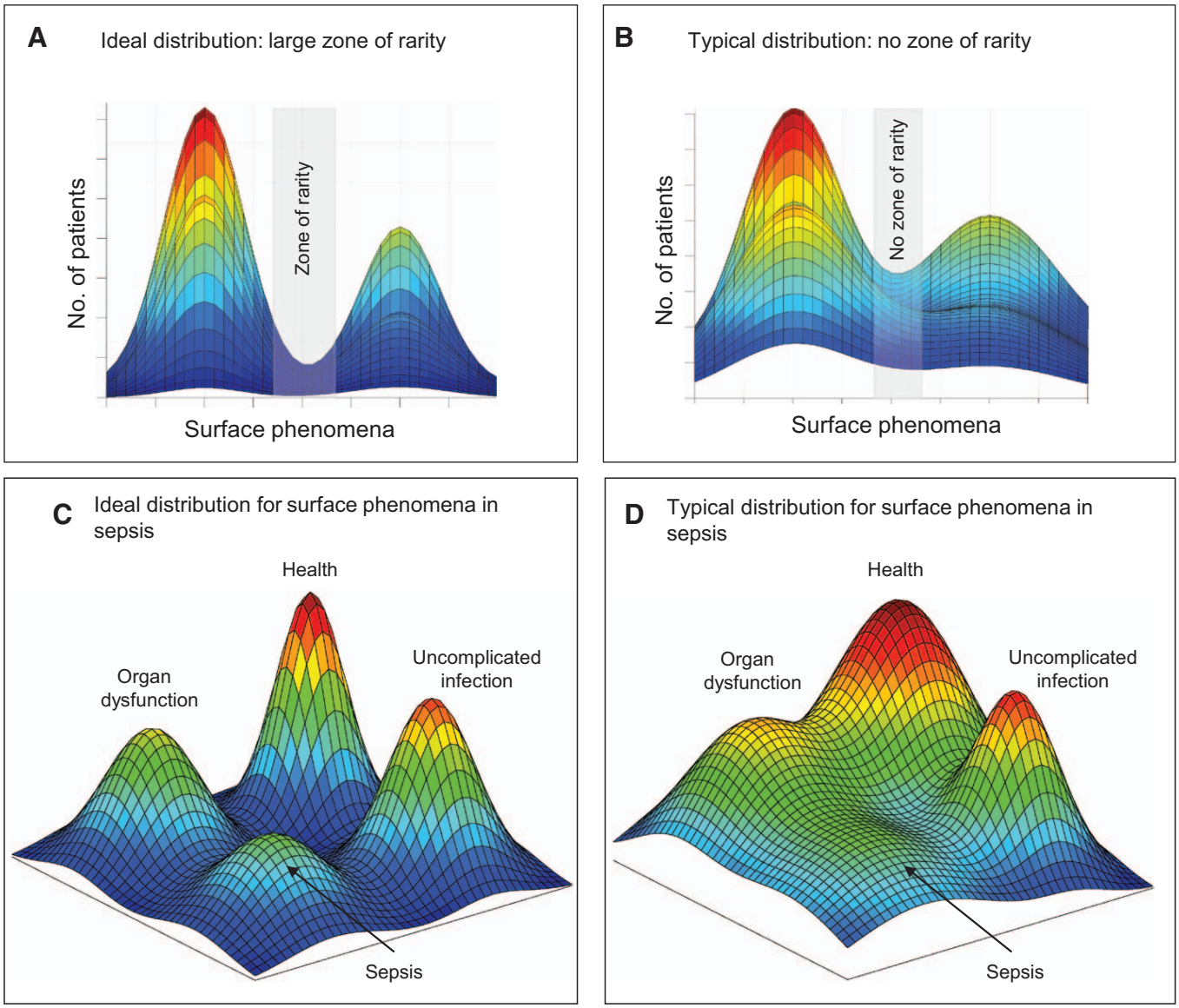

Figure 1. The "zone of rarity" problem: ideal and typical distributions of surface phenomena (clinical and biologic features) among patients with and without disease. Panels A and B illustrate situations in which a surface phenomenon (e.g., a single blood test) or set of phenomena (e.g., a combination of clinical features and blood tests) is used to separate a population into those who do and those who do not have a particular disease. Ideally (Panel A), there would be a large zone of rarity where few individuals would exhibit the test result or constellation of features at the border between health and disease. However (Panel B), most tests or combinations of tests and features are expressed on a continuum, with no zone of rarity. For example, the distribution of white blood cell count values across a population of hospitalized patients will not exhibit a zone of rarity near the upper limit of normal. Rather, many patients will have borderline-elevated values. Panel C and D show the corresponding distributions for sepsis, where surface phenomena classify patients with both infection and organ dysfunction. Although the ideal criteria (Panel C) for both infection and organ dysfunction would have clear zones of rarity, neither domains have such criteria (Panel D). For example, most organ dysfunction measures, like measures of infection, are expressed on a continuum with many patients exhibiting borderline values.

can do is to stand in the wings and help out in very minor ways (13)." For example, not well-versed in philosophy, physicians are unlikely to express opinions on the philosophic framework governing their exercise. This lack of opinion may not appear to matter initially, but becomes problematic when their deliberations lead them to difficult discussions about whether somewhat abstract concepts (e.g., dysregulated immune response) exist or not. If terms have not been defined up-front, the problems are only compounded. As poignantly noted by Scadding, even the word "disease" is "in general use without formal definition, most using it "while" allowing themselves the comfortable delusion that everyone knows what it means (14)." Indeed, Albert et al catalogued six well-held yet different views or concepts regarding what constitutes a disease (15). As discussed above, there is very little codification of any terms related to sepsis, organ dysfunction, host response, or infection.

Similarly, failure to prioritize values, or even elicit and state values, will prompt different physicians to "shout past each other" when choosing, for example, between classification schemes that rely on simple bedside criteria poorly linked to underlying

TABLE 1. Methodological Considerations for Any Disease or Syndrome Classification Exercise

Consideration	Comment	Consequences of Poor Adherence	Example of Poor Adherence When Classifying Sepsis
Epistemology	Adopt a philosophic view regarding issues such as whether unmeasurable concepts can exist or not in the framework	Disagreement over legitimacy of various concepts	No agreement over whether concepts such as sepsis, dysregulated immune response, and sepsis-induced immunosuppression, exist as clinical entities
Terminology	Define all relevant terms that frame the exercise, such as disease, syndrome, criteria, diagnosis, etc.	Disagreement over approaches and conclusions	No agreement over whether a set of surface phenomena for sepsis represent a disease, a syndrome, diagnostic criteria, diagnosis, etc.
Prioritization	Rank importance of different priorities to facilitate decision-making when faced with trade-offs	Disagreement over merits of alternative classification schemes	No agreement over whether microbiologic confirmation is necessary or not
Scientific approach			
Consensus process	Adopt explicit process for expert identification, solicitation integration, and revision of opinions	Poor reproducibility and generalizability	Different definitions from different groups
Literature review	Adopt explicit search, synthesis, and reporting methods	Poor reproducibility and generalizability	Different definitions from different groups
Empiric data analyses	Adopt explicit experimental approach, including criteria for determining performance of any diagnostic criteria (see "Usefulness," Table 2)	Poor reproducibility and generalizability	Different definitions from different groups

biologic mechanisms versus those that rely on sophisticated but expensive analyses of host and pathogen genetic, molecular, and cellular pathways involved in infection and organ function. All of these problems are faced even before specifying the technical aspects of the proposed scientific approach that would typically be described in the methods section of a medical journal article, such as choice and description of any analyzed datasets, methods to verify various biologic processes and clinical signs and symptoms, and approaches to assess agreement between different measures.

SUBSTITUTING "USEFUL" FOR ELUSIVE "GOLD-STANDARD" OR "VALID" DEFINITIONS

With the ground rules set, we return to the question of how success will be judged. The common hope is to find a gold-standard definition that stands beyond impunity. This is a search for a definition that is valid in the absolute sense, where, for example, the Shorter Oxford English Dictionary defines "valid" as "well-founded and applicable, sound and to the point, against which no objection can fairly be brought (16)." But in medicine, as in all of science, validity is a multidimensional concept, including content, construct, and criterion validity, each with further subdimensions. Any definition that scores well in one dimension and less well in another has already failed to be valid in an absolute sense, since one can bring an objection with regard to the dimension in which it fared poorly. Thus, these continuous measures of validity are part of an overall exercise determining not whether a definition is valid (a near-impossible goal) but rather whether it has utility, which, more simply can be called "usefulness (3)."

SIX DOMAINS OF USEFULNESS WHEN ASSESSING DEFINITIONS AND CLINICAL CRITERIA

We propose six domains by which the usefulness of a given definition or set of diagnostic criteria can be judged: reliability, content validity, construct validity, criterion validity, measurement burden, and timeliness (**Table 2**).

Reliability

Reliability reflects the extent to which any given measure, classification scheme, or diagnostic criterion yields stable or reproducible results. We can assess reliability in three broad ways. First, a serum lactate test is reliable if, when measured twice on the same blood sample, it returns the same, or very nearly the same, result. Generally, approved blood tests have high reliability, but reliability can be lower for the elicitation of clinical signs and symptoms and for unapproved tests, such as HLA-DR expression on monocytes. The second element is the reliability of interpretation, typically measured by interrater agreement. For example, although a chest radiograph has high test-retest reliability, its interpretation is less reliable, in that two raters may disagree regarding whether it shows pneumonia or heart failure.

TABLE 2. Six Domains of Usefulness for Potential Criteria for the Definition of Sepsis

Domain	Abbreviated Domain Definition	Illustrative Examples in Infection, Acute Organ Dysfunction, and Sepsis	
		High Performance	Low Performance
1. Reliability	Criteria yield stable reproducible results		
Test-retest	When tests are repeated	Two simultaneous serum lactates produce similar results.	Two laboratories reporting HLA-DR yield inconsistent results.
Interrater	When tests are interpreted	Two raters reading the same chest radiograph agree on presence or absence of ARDS.	Two raters of the same chest radiograph disagree on presence of ARDS
Meta-reliability	Resistance of tests or measures to changes over time or across locations that are unrelated to the disease or biology	If mechanical ventilation is required to define ARDS, then propensity to institute mechanical ventilation should be constant over time and across region.	New *International Classification of Diseases*-10 coding instructions change the way in which infection or organ dysfunction is recorded.
2. Content validity	Criteria fit with current understanding and knowledge (they make sense)	Positive cultures, vasopressor-resistant hypotension, fever, thrombocytopenia, and altered mental status fit with both clinical and biologic understanding of meningococcal sepsis.	The finding that a healthy person, when exercising, can meet some SIRS criteria, weakens the conceptual underpinning that SIRS represents a maladaptive disease state.
3. Construct validity[a]	Criteria measure what they purport to measure		
Convergent	The extent to which two or more aspects that should agree do agree	If both fever and leukocytosis are signs of infection, then they should frequently occur together (converge).	If tachycardia and fever are signs of infection, but tachycardia often occurs in the absence of fever, then convergence is low.
Discriminant	The extent to which two or more aspects that should not agree do not agree	For criteria that separate pneumonia (part of sepsis) from heart failure (not sepsis), if fever is a sign of pneumonia, and elevated JVP is a sign of heart failure, then they should not frequently occur together.	Frequent finding of fever and high JVP at the same time. Similarly, if high PCT reflects bacterial infection and high BNP reflects heart failure, concurrent elevation weakens their discriminant validity.
Multitrait multimethod matrix	Multiple assessments of agreement among different measures, comparing across traits (e.g., different infections and organ failures) and measurement methods	Highest performance is high agreement for same trait (e.g., AKI) with different methods (e.g., high creatinine and low urine output). Less valuable if 2 similar measures for the same trait agree (e.g., elevated AST and ALT for liver dysfunction).	If poor agreement between elevated lactate and hypotension, then neither alone might be considered good measures of shock.
4. Criterion validity[b]	New criteria agree with existing standard		
Concurrent	Comparison to a current standard available at the same time	Assuming positive bacterial cultures are a standard for infection, if PCT is frequently elevated in patients with positive bacterial cultures, then it has good concurrent validity	Similarly, if microbial DNA is frequently present when cultures are negative, then concurrent validity is poor.
Predictive	Comparison to a later outcome believed to be strongly associated with the disease of interest	Assuming death is more common after sepsis than after uncomplicated infection, then the finding that vasopressor requirement in patients with infection is associated with higher mortality is evidence of predictive validity for vasopressor requirement as a feature of sepsis	Among patients with presumed infection, the finding that fever is not associated with a higher likelihood of death would suggest it has poor predictive validity for sepsis.

(*Continued*)

TABLE 2. (*Continued*). Six Domains of Usefulness for Potential Criteria for the Definition of Sepsis

Domain	Abbreviated Domain Definition	Illustrative Examples in Infection, Acute Organ Dysfunction, and Sepsis	
		High Performance	Low Performance
5. Measurement burden	Burden to implement criteria		
Cost	Financial costs (to patient, provider, or healthcare system)	White blood cell count	Whole blood PCR-based gene expression microarray chip
Safety	Side effects, complications to patient	Bedside clinical examination	CT-guided tissue biopsy with radiocontrast dye
Complexity	Difficulties executing the various steps to obtain or interpret the tests and measures	Simple bedside examination	Circulating monocyte HLA-DR expression by flow cytometry
6. Timeliness	Speed with which criteria are generated with respect to the course of the disease	Simple bedside examination	Blood cultures, autopsy

HLA-DR = human leukocyte antigen D-antigen related, ARDS = acute respiratory distress syndrome, SIRS = systemic inflammatory response syndrome, JVP = jugular venous pressure, PCT = procalcitonin, BNP = B-type natriuretic peptide, AKI = acute kidney injury, AST = aspartate aminotransferase, ALT = alanine aminotransferase, PCR = polymerase chain reaction.

[a]Construct validity does not assume there is a standard measure against which others are compared. It simply searches for level of agreement or disagreement among measures. For example, if cardiac output were being measured by both Doppler ultrasound and peripheral arterial pulse contour, one would assume that neither necessarily measures cardiac output perfectly. Instead, one would assess the agreement or correlation between methods using techniques such as Bland-Altman plots. High agreement would be reassuring, even though both techniques might generate consistently biased estimates of "true" cardiac output.

[b]Because criterion validity assigns some value to the existing standard (e.g., blood cultures), one can evaluate a new measure (e.g., microbial DNA) in terms of its sensitivity and specificity for agreement with the standard. However, if one does not believe that blood cultures are in fact a good standard, then one would not necessarily assume that low sensitivity or specificity for the new measure is necessarily evidence of poor performance. For example, frequent presence of microbial DNA in the absence of positive blood cultures would suggest low specificity for blood cultures. But we may also revise our beliefs about what defines infection, and could conclude that infection is defined by the presence of microbial DNA in the blood. Under this scenario, we would revise what we considered standard and instead conclude that blood cultures have low sensitivity for microbial DNA, rather than the other way around.

The third element of reliability is the propensity to order the test, or to engage in any action that affects the measure in a way that is independent of the patient's biology. For example, if mechanical ventilation is used to define acute respiratory failure, then the number of cases of acute respiratory failure will be limited by the capacity or proclivity of the healthcare system to intubate patients. Thus, mechanical ventilation may be a reliable measure of *treated* acute respiratory failure, but not necessarily of acute respiratory failure as an isolated biologic entity, especially when used across healthcare systems of varying capacity. Similar problems arise with variable propensity to order a test or elicit and document a particular sign or symptom. This problem is not unique to sepsis and is of particular concern when interpreting epidemiologic studies, conducting audit and performance assessment, and extrapolating findings from one setting to another, such as from clinical trials to practice.

Content Validity

Content validity governs the conceptual framing of the disease or syndrome, integrating knowledge and scientific beliefs to judge the face value of a proposed definition. The deliberations of expert panels on sepsis definitions focus primarily on content validity. Thus, the 1992, 2003, and 2016 consensus conferences all retained the concepts of infection and organ dysfunction, based on their high content validity, which arises from the wealth of basic and clinical research associated with both domains. However, emerging knowledge about the complexity of the host response, coupled with awareness from both basic and clinical research that SIRS was overly simplistic, lowered the content validity for SIRS as a required causal link between infection and organ dysfunction in the definition of sepsis. High content validity is useful because it helps with acceptability but is no guarantee of truth.

Construct Validity

Construct validity is at the heart of empiric classification exercises and is defined as "the degree to which a test measures what it claims, or purports, to measure." When assessing multiple measures, one can assess the extent to which measures that should agree do agree (convergent validity) and those that should not do not (discriminant validity). For example, when examining approaches to diagnose acute left heart failure versus pneumonia, an elevated jugular venous pressure and Kerley B lines on a chest radiograph should converge with each other and diverge from (discriminate against) a history of purulent sputum or rigors and sweats. Of course, sepsis has multiple domains, relationships, and measures. Consequently, it may be wiser to use a more integrated approach, such as the multitrait-multimethod (MTMM) matrix developed in psychology (17). An MTMM matrix explores the degree of all agreement across all methods (tests or criteria) and all traits (where "trait" could be an individual organ dysfunction or an entire domain, such as organ dysfunction or sepsis). Agreement is expected to be highest when simply testing reliability (test-retest). Other comparisons are ranked based on expected agreement such

that there should be very good agreement when two measures using a similar method (monomethod) generate similar agreement when measuring the same domain (monotrait). The highest construct validity arises when two separate methods (heteromethod) that are intended to measure the same domain (monotrait) have high agreement. For example, high creatinine and anuria have high construct validity for acute renal failure.

Criterion Validity

Criterion validity assesses the extent to which a proposed measure of the disease or entity of interest agrees with an existing accepted measure that is either determined at the same time (concurrent validity) or later (predictive validity). For example, assuming positive blood cultures are an accepted measure of systemic infection, the agreement of a novel biomarker of infection, drawn at the same time (e.g., procalcitonin or a polymerase chain reaction-based assay of bacterial DNA) with positive cultures would reflect the concurrent validity of the biomarker. Similarly, assuming death following acute infection is more common following sepsis than uncomplicated infection, then agreement of any proposed criteria for sepsis among infected patients with death would be a measure of those criteria's predictive validity. It is important to set expectations when conducting criterion validity experiments. For example, sepsis is not expected to have positive blood cultures or cause death in all cases (and infected patients without sepsis may still have positive blood cultures, or die). Thus, the ceiling for these exercises will not be perfect agreement; the sensitivity and specificity should not be 100%.

Measurement Burden

Although not typically included in the evaluation of disease definitions, the burden of a given measure or set of measures is of great practical importance. The burden includes the incremental financial costs, task complexity, and clinical side effects or complications, and can be borne to varying degrees by the patient, clinician, and healthcare system. All forms of burden should be weighed from all three perspectives. Some element of burden creeps into nearly every potential measure related to sepsis. For example, measuring the Sepsis-Related Organ Failure Assessment (SOFA) score requires numerous blood tests, a clinical examination, and review of medications and organ support in the medical record (18). In a well-funded prospective clinical trial, with patient consent, it may be possible to obtain a full SOFA score every day. Assuming the costs are borne by the trial, the burden to the patient is principally that of daily venipuncture. But the costs of the blood tests and required time by staff to fully document all elements is likely too high a burden for purposes such as audits or surveillance exercises, which are more likely to rely on a simpler version of SOFA or an alternative approach to capture organ dysfunction (e.g., retrieving available information from electronic health records or hospital discharge administrative databases).

Timeliness

With many diseases, decisions are made over a period of days or weeks, which is typically long enough to await the results of most candidate measures of disease, even including complex genetic analyses of tissue biopsies. The clinical care of patients with sepsis, however, has exquisite time pressure in that both infection and life-threatening organ dysfunction generally require prompt intervention, ideally within a few hours of presentation. This time pressure creates a number of problems. For example, confirmatory evidence of an infection, such as positive blood cultures, is of little value for a prospective definition of septic shock, since the patient would likely die if awaiting treatment until cultures become available. As such, for any definition that must be made prospectively (e.g., in clinical care and clinical trials), the timeliness of any measure will be crucial.

CONCLUSIONS AND NEXT STEPS

Before examining any proposed definition for sepsis or related concepts, such as severe infection and acute vital organ dysfunction, it is helpful to set ground rules regarding underlying philosophy, terminology, and prioritization of values. These values will differ depending on whether the primary purpose is to aid clinical care, research, surveillance, or audit. Once these issues are clear, we would recommend judging any proposed criteria across six domains of usefulness. The current overarching definition for sepsis is conceptual, framing sepsis as a complex interplay of infection, host response and organ dysfunction, with no unambiguous way to measure all of these elements and their interactions. Thus, the field must deploy clinical criteria that, though intended to capture aspects of sepsis, will necessarily favor pragmatism over theory. Thus, there will likely be few criteria that perform well in all six domains. However, the good news is that the different purposes can tolerate differing performance across the domains. A set of criteria may be good for surveillance, even if not good for bedside care. The key is to match criteria whose performance across the six domains best suits the intended purpose. In the accompanying article (19), we demonstrate the ways in which these six domains of usefulness will vary in importance depending on the intended purpose.

ACKNOWLEDGMENTS

This work began through a series of discussions hosted by the Centers for Disease Control (CDC). We are extremely grateful to the CDC for their support and for the review and insightful commentary provided by colleagues at the CDC (Raymund B. Dantes, Lauren H. Epstein, Anthony Fiore, John A. Jernigan, Shelley Magill, Clifford McDonald and Daniel Pollock) and the Centers for Medicare & Medicaid Services (Megan R. Hayden, Debra C. Nichols, and Lemeneh Tefera). This work is neither a product of, nor endorsement by, either agency.

REFERENCES

1. Angus DC, van der Poll T: Severe sepsis and septic shock. *N Engl J Med* 2013; 369:840–851
2. Singer M, Deutschman CS, Seymour CW, et al; The Sepsis Definitions Task Force: The Third International Consensus Definitions for Sepsis and Septic Shock. *JAMA*, in press
3. Kendell R, Jablensky A: Distinguishing between the validity and utility of psychiatric diagnoses. *Am J Psychiatry* 2003; 160:4–12

4. Speizer FE, Ware JH: Exploring Different Phenotypes of COPD. *N Engl J Med* 2015; 373:185–186
5. Vestbo J, Rennard S: Chronic obstructive pulmonary disease biomarker(s) for disease activity needed–urgently. *Am J Respir Crit Care Med* 2010; 182:863–864
6. Esserman L, Yau C: Rethinking the Standard for Ductal Carcinoma In Situ Treatment. *JAMA Oncol* 2015; 1:881–883
7. Cohen H: Handbook of Categorization in Cognitive Science. Philadelphia, PA, Elsevier, 2005
8. Funk DJ, Parrillo JE, Kumar A: Sepsis and septic shock: A history. *Crit Care Clin* 2009; 25:83–101, viii
9. Goris RJ, Boekhorst te TP, Nuytinck JK, et al.: Does drainage of intraabdominal pus reverse. *Arch Surg* 1985; 120:1109–1115
10. Norton LW: Does drainage of intraabdominal pus reverse multiple organ failure? *Am J Surg* 1985; 149:347–350
11. Bone RC, Sibbald WJ, Sprung CL: The ACCP-SCCM consensus conference on sepsis and organ failure. *Chest* 1992; 101:1481–1483
12. Sneath PH: Some thoughts on bacterial classification. *J Gen Microbiol* 1957; 17:184–200
13. Eden M: Taxonomies of disease. *In*: The Diagnostic Process. Proceedings of a Conference Sponsored by the Biomedical Data Processing Training Program. Jacquez JA (Ed). University of Michigan, Ann Arbor, Michigan, 1963, p 57
14. Scadding JG: Diagnosis: the clinician and the computer. *Lancet* 1967; 2:877–882
15. Albert D, Munson R, Resnik M: Reasoning in Medicine. CreateSpace; 2014
16. The Shorter Oxford English Dictionary. Third Edition. Oxford, UK, Clarendon Press, 1978
17. Campbell DT, Fiske DW: Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol Bull* 1959; 56:81–105
18. Vincent JL, Moreno R, Takala J, et al: The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. *Intensive Care Med* 1996; 22:707–710
19. Seymour CW, Coopersmith CM, Deutschman CS, et al: Application of a Framework to Assess the Usefulness of Alternative Sepsis Criteria. *Crit Care Med* 2016; 44:e122–e130