NATIONAL QUALITY FORUM

Memo

TO: NQF Members

FR: NQF Staff

RE: Voting Draft Report: NQF-Endorsed Measures for Infectious Disease

DA: June 14, 2017

Background

For this project, the 16-member Infectious Disease Standing Committee met during a one-day in-person meeting to evaluate nine measures against NQF's standard evaluation criteria. The Committee evaluated four newly-submitted measures and five measures undergoing maintenance review. All nine measures were recommended for endorsement.

Comments Received

NQF solicits comments on measures undergoing review in various ways and at various times throughout the evaluation process. First, NQF solicits comments on endorsed measures on an ongoing basis through the Quality Positioning System (QPS). Second, NQF solicits member and public comments prior to the evaluation of the measures via an online tool located on the project webpage. Third, NQF opens a 30-day comment period to both members and the public after measures have been evaluated by the full committee and once a report of the proceedings has been drafted.

Pre-evaluation comments

The pre-evaluation comment period was open from February 9, 2017 to February 23, 2017 for the nine measures under review. Five pre-evaluation comments were submitted; two comments addressed the spacing of medical visits specified in the HIV Medical Visit Frequency measure; another comment suggested that a gap in HIV care should be counted at twelve months as opposed to six months. The remaining two comments addressed the sepsis management bundle measure: commenters suggested the measure was still complex and burdensome but also noted the developer's efforts to simplify the measure; commenters also asked for clarification of measure specifications. All of these pre-evaluation comments were provided to the Committee prior to their initial deliberations held during the workgroups calls.

Post-evaluation comments

The Draft Report went out for Public and Member comment April 14, 2017 to May 15, 2017. During this commenting period, NQF received nine comments from five member organizations:

Providers – 2	Professional – 1
QMRI – 1	Health Plans – 1

A complete table of comments submitted pre- and post-evaluation, along with the responses to each comment and the actions taken by the Standing Committee is posted to the project page on the NQF website, along with the measure submission forms.

Measure Specific Comments and their Disposition

Three major themes were identified in the post-evaluation comments, as follows:

1. Antibiotic Administration
2. Level of Evidence
3. Scientific Acceptability

Theme 1 – Antibiotic Administration

Measure 0500: Severe Sepsis and Septic Shock: Management Bundle received four comments concerning the administration of antibiotics in patients suspected to have sepsis or septic shock. One commenter noted that the 3-hour time window for antibiotic administration might lead to unintended consequences like antibiotic overuse. Two commenters questioned whether antibiotic administration rather than early goal directed therapy (EGDT) had a larger effect on patient survival rates as suggested by Kalil et al. (2017). One commenter suggested reducing the antibiotic administration time from three hours to two.

> **Developer Response to Comment ID 6678:** We understand that measures can have unintended consequences for patients. We remain convinced that in severe sepsis (please refer to clinical criteria used for the measure) and septic shock the risk of mortality is so high that is critical to provide early and broad antibiotic therapy. The failure to provide a proper antibiotic for a patient with severe sepsis and septic shock carries a much higher risk than the potential harm of providing a single dose of a broad spectrum antibiotic to a patient who turns out not to have severe sepsis or septic shock. This in by no means precludes the clinician's responsibility to judiciously use IV antibiotics in a way which upholds the standards of antibiotic stewardship. The 2016 Surviving Sepsis Guidelines, endorsed by the Infectious Disease Society of America speak to this question: "The rapidity of [antimicrobial] administration is central to the beneficial effect of appropriate antimicrobials. In the presence of severe sepsis or septic shock, each hour delay in administration of appropriate antimicrobials is associated with a measurable increase in mortality." When mortality already approaches 18-40% in shock states, it is unacceptable to suspend antibiotic administration pending further studies. However, in the event an antibiotic is given inappropriately in non-sepsis states, the guidelines also recommend, "Given the potential harm associated with unnecessarily prolonged antimicrobial therapy, daily assessment for de-escalation of antimicrobial therapy is recommended in patients with severe sepsis and septic shock."
>
> **Developer Response to Comment ID 6680:** The PRISM investigators have reported results in the New England Journal of Medicine (NEJM) entirely consistent with the prior Process, Promise and Arise trials, also published in NEJM. Little information is imparted by PRISM that was not already known from these previous trials. In fact, PRISM derived all data from the prior trials. We emphasize that SEP-1 is consistent with the conclusions of these trials and does not require an invasive method of patient reassessment. In regards to Dr. Kalil's publication in Critical Care Medicine, "Early Goal-Directed Therapy for Sepsis: A Novel Solution for Discordant Survival Outcomes in Clinical Trials," the major conclusion was that:
>
> > "[S]urvival discordance was not associated with differences in early goal-directed therapy bundle compliance or hemodynamic goal achievement. Our results suggest that it was associated with faster and more appropriate antibiotic co-intervention in the early goal-directed therapy arm compared with controls in the observational studies but not in the randomized trials."

While we may dispute the methods and analysis used to reach this conclusion, we again underscore that SEP-1 does not mandate an invasive reassessment but does require early IV antibiotic administration. Thus SEP-1 is consistent with the Kalil publication in its approach.

Developer Response to Comment ID 6806: The developers will take the Armstrong Institute's helpful suggestion under advisement and model data to understand how a 2 hours standard would affect the performance characteristics of the measure. We agree that earlier administration of antibiotics is the preferred approach.

Committee Response: Thank you for your comment. The Committee agrees that monitoring for unintended consequences is an important part of measure development and implementation and quality improvement programs. The Committee recommends the developers continue to modify the measure specifications (as needed) as the evidence in sepsis and septic shock continues to evolve.

Theme 2 – Level of Evidence

Measure 0500: Severe Sepsis and Septic Shock: Management Bundle received two comments noting the varying level of evidence for the different components in the measure composite. The comments suggested that the level of evidence supporting repeat lactate, fluid reassessment, and/or physical exam is not equivalent to antibiotic or IV fluid administration. The comments state that the components should not be weighted equally in the construct of the composite measure. One commenter also recommends a simplified 3-hour bundle without the repeat lactate and physical exam component.

Developer Response to Comment ID 6708: We will take under consideration the suggestion to weight elements in different ways than presently weighted. It is important to note however that as a matter of process for vetting measures, they must be advanced on the basis of accumulated data and evidence. SEP-1 continues to show a high association with reduction in mortality with the individual specified elements as documented in the submission which justified the weighting. As the submission shows, there is a large separation in mortality between those who comply with the elements in total and those who fail any one or more than one of the elements. To understand and model a proposal such as Dr. Doerfler's will require analysis of the measure as a component measure, which necessarily means analyzing the data in a fashion for which it was not designed. We will discuss with CMS Dr. Doerfler's hypothesis regarding preferential weighting of certain data elements.

Developer Response to Comment ID 6810: The developers appreciate the opportunity to respond to the High Value Healthcare Collaborative (HVHCC) which has advanced excellent work in quality improvement and care of patients with severe sepsis and septic shock. We note that the HVHCC has indicated that SEP-1 endorses an "all-or-none payment approach." SEP-1 performance does not impact payment to hospitals or providers. As regards concerns that SEP-1's composite construction does not differentiate between importance of antibiotic administration and a physical exam element (cited as skin color), the developers would like to point out that, in the current specification manual and in this NQF submission, SEP-1 does not require documentation of particular physical exam elements. A provider may now indicate simply (within the allotted timeframe) that they have "performed a physical exam" without regard to a means or method, physical exam or otherwise, and pass this data

element. In this regard, the developers would suggest that regular reassessment of a patient with septic shock as regards to perfusion status is as important as antibiotic administration and supports the composite construction as advanced in this submission. As regards the representation that "once a mature care model is in place, compliance with a 3-hr-bundle, had no impact on in-hospital, 30-day, 90-day or 1-year post discharge mortality between those receiving the full bundle vs not. This is consistent with the conclusions from the ProCESS, ARISE, and ProMISE trials," this claim is not an accurate representation of the cited trials. The 3 hour elements of care were required of every patient in the cited trials. All patients received the three hour elements (initial lactate, blood culture collection, and broad spectrum antibiotic administration) prior to randomization. Aside from this inaccuracy, it is unclear what the characteristics of a "mature care model" may be in the HVHCC's remarks, however generalizing that the 3000+ hospitals in the United States subject to SEP-1 have such a model is unsupported by any evidence to properly analyze that claim. Moreover, since HVHCC does endorse a "simplified 3-hour-bundle" it would seem to remain an important element of care. The developers do not believe that clinicians and hospitals may not define "innovative approaches to early sepsis detection" under SEP-1. One method to ascertain time zero that overrides all other methods is a provider's documentation of the time. In that regard, the HVHCC may use whatever method they prefer to set a time zero as long as their clinicians concur with the HVHCC's approach. The developers appreciate the HVHCC's suggestion to proceed with the 3 hour elements in SEP-1 and we assure them that these elements remain in this submission. As regards the representation that there is "no evidence" supporting repeat lactate assessment or a physical exam, the developers repeat that clinicians must only document reassessment of perfusion or volume status by any means of their choosing. This practice, along with repeat lactate assessment do have a supporting evidence base in the 2016 Surviving Sepsis Campaign guidelines. Reassessment is a best practice statement under the GRADE evaluation framework and repeat lactate assessment has a low quality of evidence with a weak recommendation under the same criteria. We will take under advisement that these elements should be examined further in future iterations of the specifications and would welcome the opportunity to work with the HVHCC to model these approaches. We note that the measure does not apply to critical access facilities and is not active at this time in any pay for performance programs.

Committee Response: Thank you for your comment. The Committee discussed the varying level of evidence for the different components and agreed that overall, the updated evidence is consistent with the 2016 Guidelines for Management of Sepsis and Septic Shock. Additionally, the Committee recommended that the developer complete further testing on weighting the individual components.

Theme 3 – Scientific Acceptability

Measure 0500: Severe Sepsis and Septic Shock: Management Bundle received two comments questioning the reliability and validity testing results, specifically the patient level data element percentage agreement rates. The commenters also disagreed with the guidance given to the Standing Committee on evaluating composite measures. NQF criteria states that for composite measures, validity and reliability should be empirically demonstrated at the measure score level.

Developer Response to Comment ID 6803-6804: The Federation of American Hospitals has raised many questions that were previously discussed in detail in committee. We appreciate the opportunity to summarize these issues. While burden of data collection may be greater than for other measures in healthcare, this is more than counterbalanced by severe sepsis and septic shock's' burden on the healthcare system as the number one cause of inpatient deaths in the United States and highest cost condition for hospital admissions. Evidence that the SEP-1 measure drives quality improvement was provided at NQF. The initial three quarters of data analysis show that hospitals improved their performance from quarter to quarter. In addition, the analysis revealed a statistically significant finding that there is an approximately 8.5% associated reduction in mortality in those who comply with the measure versus those who fail the measure. As measure developers we have no data to comment on the quality of responses provided by Quality Net, but we will share the feedback with the Quality Net team. We cannot address the Federation's representations regarding the motives of other agencies to utilize the measure, but we have provided substantial data that SEP-1 meets the standards set out in NQF's measure evaluation framework. As regards validity, the measure met standards for assessing validity at the performance score level, which is the proper level of evaluation for a composite measure. Additionally, it is precisely for the Federations' argument of the limited sample size (303 cases) that the data element level validity cannot serve as a valid critique of SEP-1. In addition, the measure met all reliability criteria with statistically appropriate analysis using a signal to noise methodology. While the Federation states that the element level validity testing is more important than the performance measure score testing, under the NQF measure evaluation framework, that choice is an improper standard to evaluate a composite measure. We note the Federation misinterprets the Technical Expert Panel's remarks that "the individual components may not be sufficiently reliable independently, but could contribute to the reliability of the composite performance measure." First, this quotation refers to reliability whereas the Federation was addressing validity. Secondly, the principle that the individual elements could contribute to the overall validity at the level of the performance score is precisely the point of the Technical Expert Panel's comment. This rationale is why element validity is not the criterion for composite measures. Finally, the Federation has advanced no evidence to evaluate the claim that the measure does not meet the validity criteria set out by NQF at the performance measure score level.

The Federation has inadvertently misstated the evaluation criteria. Specifically, subcriterion 2d states that "[f]or composite performance measures, empirical analyses support the composite construction approach…" Nothing in the Federation's remarks indicates that the composite construction approach is under question. As regards "missing data," the measure evaluation frame work considers this under subcriterion 2b7, which requires that the measure "analyses and identifies the extent and distribution of missing data (or non-response) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders and how the specified handling of missing data minimizes biases." This requirement is different than the scenario described by the Federation which equates the variation in element level validity testing with missing data. This equivalency is incorrect insofar as the analysis of the complete data shows that some degree of variation actually exists – if the data were "missing" a showing of variation could not be made. In this regard, the developers stand by their submission that all data is present and cannot be missing as part of the reporting requirements met by 99.9% of over 3200 acute care hospitals in three consecutive quarters of analyzed data.

To the extent that SEP-1 data elements may be contained in an electronic health record, the developers will take the Federation's excellent suggestion under advisement to consider, if feasible, the measure as an e-measure. The Federation implies that hospitals do not understand why they fail the measure given its composite nature. The developers believe, however, that by analyzing the point of failure in the measure framework, providers know exactly where weak spots are in their clinical care processes. For example, failure of the antibiotic element indicates a process in antibiotic administration that needs attention. In addition, the software provided by major vendors to hospitals to report the measure specifically categorizes the level of the fallout for facilities. The developers strenuously object to the characterization that the measure has limited value in improving patient care: in over 600,000 patients captured by the measure there has been an approximately 8.5% reduction in mortality in those compliant with the measure versus those who were not. For the population of submitted cases, this represents a potential lives saved calculation of over 7,500 patients in the first three quarters of data for the measure.

Developer Response to Comment ID 6811-6813: The developer's appreciate the opportunity to respond to the remarks of the American Medical Association (AMA) and clarify the operation of SEP-1. As regards to Dr. Pronovost's publication in the American Journal of Medical Quality regarding possible unintended consequences of quality measures, we note that Dr. Pronovost is the Director of the Armstrong Institute for Patient Safety and Quality which has recommended that SEP-1 be re-endorsed with the exception of tightening the antibiotic administration requirement from 3 hours to 2 hours. Please see the submitted comment of Dr. Matt Austin, PHD, on behalf of the Armstrong Institute for Patient Safety and Quality at Johns Hopkins University, which was also received during the post-evaluation comment period.

With respect to the specific concern that fluid administration as specified in SEP-1 may be harmful to patients with left ventricular systolic dysfunction, the AMA cites an opinion article authored by an emergency medicine resident. We note that this opinion is not representative of any clinical trial, observational or randomized, controlled or not. The opinion cites another article by Boyd 2011 which indicates that a positive fluid balance and elevated CVP are associated with increased mortality. The developers note that SEP-1 does not advocate for a "positive fluid balance" which refers to volume status over several days of care. SEP-1 is limited to initial resuscitation and the first 6 hours of care after presentation. In addition, SEP-1 does not advocate for CVP measurement, and certainly not an "elevated CVP." Another citation in the resident's article is Pudilo 2012 which reports frequency of myocardial dysfunction in severe sepsis and septic shock. The article by Pudilo actually points to a reason for proper volume resuscitation of patients: the well-known presence of severe sepsis induced myocardial dysfunction. The salient finding is not that the patients have intrinsic heart disease, but rather that sepsis has caused impaired myocardial function. In addition, Pudilo does not conclude that a 30 ml/kg initial fluid bolus as an initial resuscitation strategy in septic shock is detrimental to patients. This Pudilo reference does not support any of AMA's criticism of the SEP-1 measure.

On the broader concern about the potential risks of fluid resuscitation, we note that there is no published evidence from any randomized controlled trial which indicates that patients with severe sepsis and known congestive heart failure or renal failure who receive a fluid bolus for initial resuscitation do worse than other sepsis patients in terms of mortality. In fact, even the examination of the large trials on septic shock do not support this contention (EGDT 2001, Process 2014, Promise 2014, Arise 2015). In fact the only published evidence on the topic concludes that for patients with intermediate lactate vales of 2-4 mmol/L who receive the full fluid bolus of 30 ml/kg with congestive

heart failure and renal failure have lower mortality than their counterparts without these co-morbidities. (See: Lui V et al. Fluid Volume, Lactate Values, and Mortality in Sepsis Patients with Intermediate Lactate Values. Ann Am Thorac Soc Vol 10, No 5, pp 466-473, Oct 2013).

The thrust of the AMA's comments on this topic regarding LVSD is that physician judgment should be preserved. The developers agree with the AMA that physician judgment is paramount and agree that providers should exercise their best judgment informed by the evidence when caring for sepsis patients. SEP-1 is not a prescriptive recipe for all patients with severe sepsis and septic shock, but rather a measurement strategy for processes of sepsis care. The developers fully expect that practitioners will do what is best in their understanding for each patient, which may result in deviation from SEP-1. Ultimately when sufficient data is amassed, best compliance, which takes into account those necessary deviations, will be known. In that regard, there is no expectation or goal of a 100% compliance with SEP-1. However, we emphasize that all current evidence suggests better mortality with higher compliance.

As regards to the concern that a precisely specified measure should account for how unplanned drug shortages could impact an individual hospital's performance and the concern that mortality varied with a shortage of nor-epinephrine, we note that SEP-does not require the use of any one particular vasopressor. We also note that SEP-1 is a process measure, not an outcome measure such as mortality. Finally, we note that it is likely that all hospitals would be affected by any shortage in a short period of time.

Turning to Dr. Kalil's publication in Critical Care Medicine, "Early Goal-Directed Therapy for Sepsis: A Novel Solution for Discordant Survival Outcomes in Clinical Trials," the major conclusion was that "[s]urvival discordance was not associated with differences in early goal-directed therapy bundle compliance or hemodynamic goal achievement. Our results suggest that it was associated with faster and more appropriate antibiotic co-intervention in the early goal-directed therapy arm compared with controls in the observational studies but not in the randomized trials." While we may dispute the methods and analysis used to reach this conclusion, we again underscore that SEP-1 does not mandate Early Goal Directed Therapy and SEP-1 does require early antibiotic administration. Thus SEP-1 is consistent with the Kalil publication in its approach. The PRISM investigators have reported results in the New England Journal of Medicine (NEJM) entirely consistent with the Process, Promise and Arise trials, also published in NEJM. Little information is imparted by PRISM that was not clearly know from the three other trials, and in fact PRISM derived all data from the prior trials. In any case, SEP-1 as specified is consistent with the conclusions of these trials.

AMA has stated concerns with SEP-1's measure performance characteristics. For its validity, the measure met standards for assessing validity at the performance score level, which is the proper level of evaluation for a composite measure. In addition, the measure met reliability criteria with statistically appropriate analysis using a signal to noise methodology. While AMA represents that element level validity testing is more important than the performance measure score testing, this is an improper standard to evaluate a composite measure under the NQF measure evaluation framework. Of substantial importance, we note that AMA misinterprets the Technical Expert Panel's remarks that "the individual components may not be sufficiently reliable independently, but could contribute to the reliability of the composite performance measure." First, the quoted principle that the individual elements could contribute to the overall validity at the level of the performance score is precisely the point of the Technical Expert Panel's recommendation not to focus on element level testing for a composite metric. Second, the AMA has advanced no evidence to evaluate the claim that the measure

does not meet the validity criteria set out by NQF at the performance measure score level. In the evaluation of a metric, it would be improper to move the goal post mid-evaluation.

Regarding their other comments on validity, the AMA incorrectly cites subcriterion 2d. Subcriterion 2d states that "[f]or composite performance measures, empirical analyses support the composite construction approach…" Nothing in the AMA's remarks indicates that the composite construction approach is under question. In regard to "missing data," the measure evaluation frame work actually considers this under subcriterion 2b7, which requires that the measure "analyzes and identifies the extent and distribution of missing data (or non-response) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and non-responders and how the specified handling of missing data minimizes biases." This requirement is different than the scenario described by AMA which equates the variation in element level validity testing with missing data. This equivalency is incorrect insofar as the analysis of the complete data shows that some degree of variation actually exists – if the data were "missing" a showing of variation could not be made. In this regard, the developers stand by their submission that all data is present and cannot be missing as part of the reporting requirements met by 99.9% of 3225 acute care hospitals in three consecutive quarters of analyzed data.

In summary, the developers appreciate the AMA's comments and suggestions. We agree with the AMA that this disease which places an unacceptable burden in terms of deaths and expenditures on the healthcare system deserves a rigorous and robust measure. For this reason we are proud that SEP-1 (which has over 600,000 reported cases since inception) has an associated 8.5% reduction in mortality for those cases that were compliant with the measure versus those which were not.

We look forward to tracking the ongoing impact of the SEP-1 measure on sepsis quality improvement and will share our findings with all concerned stakeholders as they become available.

NQF Response: Thank you for your comment. NQF composite performance measure evaluation guidance (2013) states that validity testing is directed toward the inferences that can be made about accountable entities on the basis of their performance measure scores. For the purposes of endorsing composite performance measures, validity testing of the constructed composite performance measure score is more important than validity testing of the component measures. Even if the individual component measures are valid, the aggregation and weighting rules for constructing the composite could result in a score that is not a true reflection of quality (p. 12-13). Thus, the data element testing results can be considered, but the score-level results are of greater interest. Please note that the statement in NQF's composite report that states "the individual components may not be sufficiently reliable independently, but could contribute to the reliability of the composite performance measure" is actually referring to score-level reliability testing of the components, not data element-level testing of the components. In other words, it is possible that there is not sufficient signal for an individual component, but inclusion of that component will increase the reliability of the overall composite. The updated reliability and validity composite score level testing provided by the developer meet NQF criteria for composite measures as indicated during the in-person meeting and in the draft report.

Committee Response: Thank you for your comment. The Committee encourages the developers and CMS to provide education and outreach, continued efforts to decrease abstraction burden and successive data element validation testing.

NQF Member Voting

Information for electronic voting has been sent to NQF Member organization's primary contacts. Accompanying comments must be submitted via the online voting tool.

Please note that voting concludes on June 28, 2017 at 6:00 PM ET –no exceptions.