Addressing Low Case-Volume in Healthcare Performance Measurement of Rural Providers

RECOMMENDATIONS FROM THE MAP RURAL HEALTH TECHNICAL EXPERT PANEL

FINAL REPORT

MARCH 29, 2019

NATIONAL QUALITY FORUM

This report is funded by the Department of Health and Human Services under contract HHHSM-500-2017-00060I Task Order 75FCMC18F0004.

CONTENTS

EXECUTIVE SUMMARY

Low case-volume presents a significant measurement challenge for many rural providers, particularly when they want to compare their performance to that of other providers (both rural and nonrural), identify topics for improvement, or assess change in quality over time. Rural areas are, by definition, sparsely populated, and this can affect the number of patients eligible for inclusion in healthcare performance measures, particularly condition- or procedure-specific measures. Other challenges faced by rural residents, such as distance to care or lack of transportation, can also lead to low case-volume in healthcare performance measurement. In 2018, as an extension of NQF's work in convening the MAP Rural Health Workgroup, CMS tasked NQF with eliciting expert input on promising statistical approaches that could address the low case-volume challenge, as it pertains to healthcare performance measurement of rural providers.

To develop recommendations to address the low case-volume challenge for rural providers, NQF convened a five-member **Technical Expert Panel** (TEP) of statistical experts and measure methodologists. As part of the effort, the TEP reviewed previously identified approaches to the low case-volume challenge and offered new recommendations as appropriate. In fulfilling its charge, the TEP considered exemptions for reporting requirements for rural providers in various CMS quality programs, as well as the heterogeneity of both the residents and healthcare providers in rural areas.

The TEP's four key recommendations to address the low case-volume challenge are to:

- "Borrow strength" for low case-volume rural providers to the extent possible by systematically incorporating additional data as needed (e.g., from past performance, from other providers, from other measures, etc.)
- Recognize the need for robust statistical expertise and computational power to implement the recommended modeling approach of borrowing strength
- Report exceedance probabilities (exceedance probabilities, like confidence intervals, reflect the uncertainty of measure results)
- Actively anticipate the potential for unintended consequences of measurement

TEP members also suggested several additional ideas for future work that could further address the low case-volume challenge for rural providers.

Although CMS is the primary audience for the recommendations in this report, many other stakeholders in the healthcare performance measurement enterprise also can benefit, given the actionability of the approaches and activities that the TEP included in its recommendations.

The TEP's recommendations advance the science of performance measurement by addressing, in concrete, sophisticated, and yet understandable ways, the continuing and important challenge of low case-volume, particularly for rural providers.

Perspective of a Rural Resident

Mary, a 57-year-old physical education teacher in northern Michigan, needs knee replacement surgery. She is trying to decide whether to have this procedure done at rural hospital A or urban hospital B. Through an online search, she discovers hospital-level reports for a patient-reported outcome measure that evaluates postoperative patient knee pain and stiffness. However, while the result for this measure is shown for hospital B, no values are shown for hospital A. At first, Mary concludes that hospital A provides poor quality care. Upon further investigation, she learns that not enough patients receive knee replacement surgery at hospital A to report values for this particular measure. Thus, the lack of measure results provides some information on the volume of procedures at hospital A. Yet because sufficient data for hospital A are not available, Mary cannot take into account the quality of care provided at that hospital for patients having knee replacement surgery thus cannot make a fully informed choice between hospital A and hospital B.

Mary faces a conundrum common for many publicly reported measures. Rural providers do not always meet the minimum number of cases required to report reliable values. Recommendations made by the MAP Rural Health TEP may allow for more complete reporting of measure results for rural providers, ultimately enabling patients like Mary to make decisions about where to receive care across a range of hospitals.

Provider Perspective

Dr. Q, a physician based in central Kansas, is known locally for her outstanding patient care and is confident in her ability to provide sound treatment. However, because Dr. Q sees relatively few patients for particular conditions, her scores for several quality measures are not publicly reported. Dr. Q is an excellent care provider, yet she cannot demonstrate this due to the low case-volume issue. As a result, rural patients will not be able to find physicians like Dr. Q, when searching for the best available provider. Furthermore, the lack of reliable measure results also may render Dr. Q ineligible for pay-for-performance programs or prevent her from receiving incentive payments from such programs.

Dr. Q, like many rural providers, is at a disadvantage relative to providers with higher case volumes, because her low case-volume precludes reliable estimation and reporting of key performance metrics. Several recommendations of the MAP Rural Health TEP address this issue specifically, and if implemented, will help ensure reliable performance assessment and reporting across a range of facilities in both urban and rural settings across the United States.

INTRODUCTION

Low case-volume presents a significant measurement challenge for many rural providers, as it affects the reliability, validity, and utility of many measures that might otherwise be available to them for assessing care quality. In 2014, the Department of Health and Human Services (HHS) funded the National Quality Forum (NQF) to convene a multistakeholder Committee to identify challenges in healthcare performance measurement for rural providers[a] and to make recommendations for meeting these challenges.[1] The low case-volume challenge was a key focus of that effort and accordingly, several of the Committee's recommendations addressed this challenge directly.

That Committee also recommended that CMS create a Measure Applications Partnership (MAP) Workgroup to provide advice on the selection of rural-relevant measures for quality improvement programs. MAP is a public-private partnership of healthcare stakeholders, convened by NQF, that provides input to the Department of Health and Human Services (HHS) on the selection and alignment of performance measures for public reporting and performance-based payment programs. In 2017, recognizing the MAP's lack of representation from rural stakeholders, CMS tasked NQF to establish a MAP Rural Health Workgroup. This Workgroup provides a rural perspective on various issues pertaining to healthcare performance measurement to CMS, as well as to the other MAP workgroups and the MAP Coordinating Committee. To date, the MAP Rural Health Workgroup has identified a core set of the best available rural-relevant measures to address the needs of the rural population and provided recommendations, from a rural perspective, regarding measuring and improving access to care.[2] The Workgroup also offered a rural perspective to the MAP Clinician Workgroup during its **2018 pre-rulemaking activities**.

In 2018, as an extension of NQF's work with the MAP Rural Health Workgroup, CMS tasked NQF with eliciting expert input on promising statistical approaches that could address the low case-volume challenge. To accomplish this task, NQF launched a national call for nominations and convened a five-member **Technical Expert Panel** (TEP) of statistical experts and measure methodologists. Members of this MAP Rural Health TEP are proficient in Bayesian statistics, small area estimation,[b,3] nonparametric statistics, and performance measure development, and they have hands-on experience in quality measure reporting for low-volume rural providers.

NQF charged the TEP with reviewing previously identified approaches to the low case-volume challenge and offering new recommendations as appropriate. In fulfilling this charge, the TEP considered exemptions for reporting requirements for rural providers in various CMS quality programs, as well as the heterogeneity of both the residents and healthcare providers in rural areas.

As the sponsor of this effort, CMS is the primary audience for the recommendations in this report. However, the TEP's recommendations include approaches and activities that are actionable for other public and private stakeholders in the healthcare performance measurement enterprise. These include measure developers, those who implement and report measures in both public and private quality improvement and accountability programs, entities that fund

a In this report, "providers" are defined broadly as those entities that can be held accountable for the provision of healthcare. Thus, "providers" include individual clinicians, clinician groups, hospitals, post-acute care settings such as nursing homes, and other entities such as health plans, health systems, states, and regions, as well as programs such as Medicaid.

b In statistics, estimation for domains for which there are too little data for the usual direct estimates to work (e.g., geographical areas, healthcare providers, etc.) is known as small-area estimation.

or otherwise sponsor performance measure development or implementation, policymakers, quality improvement professionals, and healthcare providers.

The remainder of this report is organized into five sections that provide context for this effort and describe the deliberations and recommendations of the TEP. The first section describes the low case-volume challenge for rural healthcare providers, provides an overview of the implications of low case-volume on the reliability and validity of measurement, and summarizes the decisions of the TEP in terms of how it considered low case-volume for the purposes of this report. The following section reviews previously articulated solutions to the low case-volume challenge and provides the TEP's input on the strengths and weaknesses of those approaches. Building on its discussions of previously identified solutions, the next section presents the TEP's four key recommendations to address the low case-volume challenge in healthcare performance measurement for rural providers. The next section includes several additional recommendations by the TEP for future activities and research. The final section summarizes the TEP's recommendations and describes how they advance the field of healthcare performance measurement. Three appendices provide additional details relevant to this work. **Appendix A** lists the TEP members and NQF staff involved in this effort. **Appendix B** provides additional detail on NQF's approach and timeline for the work described in this report. **Appendix C** includes the comments submitted by NQF members and the public in response to a draft version of this report.

BACKGROUND

Approximately 19 percent of the U.S. population—more than 59 million individuals—live in rural areas.[4] Rural residents are more disadvantaged than those in other areas, particularly with respect to sociodemographic factors, health status and behaviors, and access to the healthcare delivery system.[4–10] Moreover, rural healthcare providers face many challenges in reporting quality measurement data and implementing care improvement efforts to address the needs of their populations. Low case-volume is a key challenge of healthcare performance measurement for rural providers, particularly when they want to compare their performance to that of other providers (both rural and nonrural), identify topics for improvement, or assess change in quality over time. Often, rural physicians or facilities have too few patients who meet inclusion criteria for healthcare performance measures, particularly those that are condition- or procedure-specific. Other challenges faced by rural residents, such as distance to care or lack of transportation, can also lead to low case-volume for performance measurement if these residents forgo care due to such challenges.

Prior to offering specific recommendations to address the low case-volume challenge, the TEP discussed the ways in which low case-volume impacts the reliability, validity, and usability of healthcare performance measures and the ways that low case-volume might be considered for the purposes of this report.

Impact of Low Case-Volume on the Reliability and Validity of Measurement

The level of confidence in the conclusions about quality is directly related to the reliability and validity of measurement. The concepts of both reliability and validity can be applied to the individual data elements that are used in a measure and to the computed performance measure score. NQF currently defines reliability as the repeatability or precision of measurement and validity as the correctness of measurement.[11] Reliability of data elements refers to repeatability and reproducibility of the data elements for the same population in the same time period. Reliability of the measure score refers to the proportion of variation in the performance scores due to systematic differences across the measured entities (signal) in relation to random variation (noise). Validity of data elements refers to the correctness of the data elements as compared to an authoritative source. Validity of the measure score refers to the correctness of conclusions about quality that can be made based on the measure scores.

The reliability of the measure score is a function of sample size, the magnitude of the between-provider variance, and measurement error. Providers with smaller sample sizes (i.e., low case-volume) are likely to have lower score-level reliability estimates. Low case-volume can also affect the validity of measurement, although the relationship is not as clear as it is with reliability. For example, for measures that are risk-adjusted, low case-volume could affect the development of the risk-adjustment approach, which itself is one aspect of validity. After reviewing the above definitions, the TEP agreed that low case-volume is of concern primarily in terms of its impact on the reliability of the measure score, and the majority of its recommendations address this concern.

Defining Low Case-Volume

TEP members considered the following ways of defining low case-volume for the purposes of this report:

- Too few individuals meet the measure denominator

- Too few individuals meet the measure numerator
- As defined by specific program reporting requirements (i.e., reporting thresholds)

After reviewing each of these possible definitions, the TEP agreed to consider low-case volume primarily as having too few individuals that meet the measure denominator criteria. Members noted that some measures, by design, will have very low numerator counts (e.g., measures of patient safety "never events"), and that consideration of the magnitude of the numerator, relative to that of the denominator, may be of more interest than focusing on the numerator. Regarding using specific program reporting requirements to define low case-volume, TEP members noted that thresholds for reporting often are implemented due to concerns about privacy, which are different from concerns regarding low case-volume and its resulting effects on score-level reliability. Thus, the TEP decided to consider the various program-specific thresholds on a case-by-case basis, if necessary, rather than use them to define low case-volume for this report.

The TEP also discussed whether to consider complete lack of service provision (e.g., a hospital does not perform deliveries) as a part of their deliberations. Members agreed that this is a missing data problem within the context of composite measures and program design, rather than a low case-volume problem. Therefore, they decided that this situation is out of scope for this report.

Content and Use of the TEP's Recommendations

TEP members agreed that they cannot make specific measure development recommendations (e.g., providing the specifications for a single statistical model), given the number and types of measures that could be developed for use in rural settings, as well as the varying goals of programs that would use such measures. Instead, they agreed that their recommendations should provide general methodological guidance to the field.

The TEP emphasized two general principles that apply to healthcare performance measurement when there is a lack of data (e.g., when rural providers have too few patients for reliable measurement). First, the "usual" estimators or approaches that would otherwise be used in a data-rich environment are either not possible or are not useful. Second, when using alternative approaches, such as those recommended by the TEP, the resulting estimates—while better than nothing—are not an ideal substitute for what would be used if data were available. Consequently, when using these alternative approaches, the comparisons and interpretations that might be possible for estimates derived in a data-rich environment may not be appropriate. Throughout its deliberations, the TEP cautioned that the suitability of a particular measure development approach must be evaluated against the intended purpose of the program in which it will be used. Conversely, when selecting measures for a particular program or use, the approaches used when developing the measures, as well as the resulting strengths and limitations of those approaches, must be aligned with the program purpose.

TEP CONSIDERATION OF PREVIOUSLY IDENTIFIED POTENTIAL SOLUTIONS TO THE LOW CASE-VOLUME CHALLENGE

As noted earlier, the low case-volume challenge was a major area of focus for NQF's 2015 Rural Health Committee. Building on the approaches identified through an environmental scan conducted to inform its discussions, that Committee proposed several additional potential solutions to address the low case-volume challenge for rural providers. The majority of the 13 proposed solutions identified through the environmental scan and by the 2015 Committee address various aspects of measure specifications. However, two of the proposed solutions address how measure results are reported.

As a first step in forming its recommendations for this project, the TEP considered all of the previously proposed solutions and commented on their strengths and weaknesses, as described below. Beyond serving as a useful way to acquaint the TEP with previous deliberations on the topic of low case-volume and healthcare performance measurement for rural providers, the results of this activity provide a valuable supplement to the work of the 2015 Committee.

Potential Solutions that Address Measure Specifications

Eleven of the 13 previously articulated potential solutions to the low case-volume challenge address how measures are constructed. These solutions can be categorized in terms of recommendations about the target population for the measure, the calculation used to represent the measure focus, the data that are included in the measure, and the mechanics of the measure calculation.

Recommendations Regarding the Target Population

A measure's target population refers to those individuals who are broadly captured by the measure (e.g., patients with diabetes who are discharged from a hospital, or enrollees in a health plan). The first three of the following recommendations reflect the 2015 Rural Health Committee's charge to address the challenge of low case-volume for rural providers, particularly in the context of selection of measures for CMS pay-for-performance programs. Three of the four recommendations address the low case-volume challenge by promoting use of measures with large target populations, while the fourth promotes specification of measures that capture the target population to the extent possible by reconsidering measure exclusions.

Select Measures that are Broadly Applicable across Rural Providers

The 2015 Rural Health Committee identified several topic areas (e.g., vaccinations, screening, blood pressure control, diabetes control, medication reconciliation) that apply to a large proportion of patients served by rural providers. Its members recommended use of measures addressing these topics when assessing the quality of care by rural providers in pay-for-performance programs.

The TEP agreed that this approach would reduce the low case-volume problem to a large extent, as well as help to maximize statistical power. This approach can be easily combined with other criteria that are deemed important when deciding on measures to include in particular programs.

From the rural perspective, the TEP agreed that measures that assess screening, immunizations, diabetes, and transitions of care are not only

broadly applicable to large numbers of patients but are particularly relevant to rural populations. Selection of such measures would likely increase the number of rural providers who could participate meaningfully in various types of accountability programs. This approach also could help to reduce measurement burden for rural providers.

The TEP also noted some potential drawbacks of this approach. In particular, limiting the selection of measures to those that are applicable for most rural providers places artificial constraints on the available measures. This could result in the neglect of other measures that are important for rural populations. For example, a focus on screening or immunizations might jeopardize quality improvement efforts in rural areas for other important conditions or healthcare activities such as specialty care or surgical services. TEP members also suggested that such a focus might, in some cases, tilt selection away from use of outcome measures. Finally, there may not be an objective way to determine which measures meet the criterion of "broadly applicable" (or a way to otherwise reach consensus on what it means to be broadly applicable).

Consider Measures that Reflect the Wellness of the Community

Measures that reflect the wellness of the community (a particular subset of population-based measures such as admission rates for long-term consequences of diabetes or percentage of low birthweight births) typically are constructed such that the accountable entity is either a particular geopolitical area (e.g., a state or community) or some other subpopulation that is independently assessed within the context of a larger geopolitical area (e.g., a state Medicaid program). Again, the impetus for this recommendation by the 2015 Rural Health Committee was to provide suggestions for the types of measures best suited for assessing quality of care for rural residents. That Committee acknowledged the importance of population-based measures and agreed on their utility for addressing the low case-volume challenge. Its members also recognized the potential for such measures to be used for internal quality improvement at the provider level. They did not, however, support the use of such measures in pay-for-performance initiatives for rural providers, in part due to differences between communities in terms of culture, availability of resources, and feasibility of data collection.

The TEP agreed that use of population-based measures of community wellness could serve as a potential solution to the low case-volume problem, due to the large number of individuals who would be included in such measures. However, the TEP noted several limitations of this approach, including the subjectivity in defining "community" and the fact that the health of a community is determined only in part by healthcare. TEP members also agreed that population-based measures, overall, are less useful than other types of measures in guiding improvement activities or other types of decision making (e.g., consumer choice), as such measures do not assess the care that is delivered by a specific provider (i.e., clinician, hospital, etc.). Finally, such measures do not allow for accountability at an individual provider level.

Reconsider Exclusions for Existing Measures

Although acknowledging the need to exclude certain individuals from particular measures, the 2015 Committee recommended that measure developers explicitly consider the impact of low case-volume when specifying measure exclusions (i.e., individuals or events that are part of the measure's target population but are not captured by the measure). The Committee cited the Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) measures as an example. These measures exclude patients who reside in nursing facilities or who receive hospice care, due to the difficulty in collecting data from these patients and the concern that they may conflate their hospital experiences with those of the nursing facility or hospice.

The TEP agreed that explicitly considering the impact of low case-volume when specifying measure exclusions could help to address the low case-volume challenge. However, TEP members noted that if exclusions are used as a way to ensure patient or treatment homogeneity, finding a reasonable workaround likely would increase the complexity of the measure development process and, potentially, the measure specifications. From the rural perspective, the TEP cautioned measure developers to consider any disproportionate effects that deleting or otherwise modifying exclusions may have on rural providers, particularly if done primarily as a way to increase the denominator size. For example, some hospital readmission measures exclude patients who are transferred from one hospital to another. While this decision decreases the number of patients included in the measure denominator, it provides a de facto "reward" for rural providers who do the right thing when transferring patients, by not holding them accountable for a potential readmission.

Develop Composite Measures that Expand the Number of Patients Captured by Measurement

Composite measures typically combine information from multiple individual performance measures into a single measure that results in one overall score.[c] Such composite measures can be constructed using both "homogeneous" measures (those with similar, overlapping information) as well as "heterogeneous" measures (those from different domains). The 2015 Rural Health Committee focused their discussion of composite measures on the components of composite measures. Specifically, members of that Committee agreed that composite measures should include individual component measures that are applicable across rural providers (i.e., those that are not susceptible to low case-volume).

The TEP agreed that composite measures have the potential to address the low case-volume challenge. When variation in individual component measures "cancels" each other out (such as when the component measures are weakly correlated at the patient level), composite measures may have reduced sampling variations, and hence, improved stability. Use of composite measures can also effectively reduce the number of measures that are reported (although their use likely would not reduce the burden of data collection). Because composite measures, by definition, have multiple components, they may more effectively capture the complexities of care for certain healthcare conditions. They also may be more informative to patients and consumers in reflecting the quality of care for such conditions by providing a single, summary estimate that they can more readily understand.

However, the TEP highlighted the practical and statistical challenges of choosing the components that are included in composite measures and determining the relative weights of the components. In addition, because composite measures comprise multiple measures of care quality, a mid-range score provides little information to providers that will help them identify opportunities for improvement (unless scores for individual components also are provided). Lastly, the quality signal[d] (or the changes in the quality signal) in one or a few component measures can be masked by measurement error associated with other component measures.

Recommendations Regarding the Calculation for the Measure Focus

Three of the 13 previously proposed solutions to the low case-volume challenge suggest alternative ways to specify the measure focus. The "measure focus" is a term used to describe, conceptually,

c Examples of such composite measures include NQF **0531 Patient Safety and Adverse Events Composite** and NQF **3030 STS Individual Surgeon Composite Measure for Adult Cardiac Surgery**.

d The "quality signal" represents the ability to distinguish true differences in performance among providers (the signal) from measurement error (the noise).

what the measure is about or what is being measured. Usually there are several ways that a particular measure can be constructed to address the measure focus. For example, a measure of mortality could assess the total number of deaths, the average number of deaths, the rate of death per some population or timeframe, etc. The three alternative approaches for specifying the measure focus (using continuous variables, ratio measures, and measures that do not have a denominator), discussed below, currently are used infrequently in healthcare performance measurement programs.

Measures Constructed Using Continuous Variables

The 2015 Rural Health Committee suggested that measuring an aspect of care by using a continuous variable to detect meaningful differences between providers may require a smaller sample size than what would be needed if the measure used a dichotomous variable (e.g., assessing the time until a medication is given rather than just whether or not a medication was given). However, its members cautioned about use of continuous measures for rural providers because such measures may be particularly sensitive to outliers.

The TEP agreed that measures that yield continuous results can be statistically more powerful than those that are constructed to result in categorical (typically binary) values. On the negative side, however, the TEP noted that continuous measures are not as easily interpretable and may not be as easily or understandably presented as their categorical counterparts. For example, results of a measure assessing the time until a medication is given likely would need to be categorized with specific time intervals or periods (e.g., within 6 hours, 10 hours, or 24 hours) to facilitate reporting and understanding of the results. However, determining the appropriate categorization may be challenging. Moreover, patients or other consumers who might be using the results for decision making may not understand the implications of clinically meaningful categorizations.

Ratio Measures

Ratio measures are those in which the numerator is not necessarily a part of the denominator. For example, one could construct a ratio measure that assesses prevalence of bloodstream infections where the numerator sums the number of bloodstream infections and the denominator sums the number of days during which patients have a central line. The 2015 Rural Health Committee noted that such measures could help to circumvent the low case-volume problem because each patient could contribute many "units" to the denominator.

The TEP agreed that such measures could address the low case-volume challenge because the size of the denominator would increase. TEP members also thought such measures may be more informative than typical binary measures. However, they noted that care must be taken when specifying the denominator to be sure that it is not related to care quality (e.g., in the above example, the concern is that the number of days on a central line may, itself, be an indicator of (low) quality). They also suggested that although this form of measurement would have a larger denominator, it may not be adding meaningful information, and therefore would not actually increase the reliability of the measure.

Measures that Do Not Have a Denominator

One potential solution to the low case-volume challenge that was identified through the 2015 environmental scan but not specially discussed by the Rural Health Committee is to calculate measures without specifying a denominator (e.g., number of infections per month or time since last adverse event).

The TEP agreed that such measures would provide one less source of sampling variation (i.e., in the denominator) and also agreed that some of these measures (e.g., measures of "never events") could potentially be used in accountability applications. However, the TEP acknowledged that such measures would still be volume-dependent. TEP members also recognized that for many scenarios,

measures of proportions or other types of ratios may be especially informative or useful for certain topic areas and therefore cannot be replaced by measures that include only numerators. Using "numerator-only" measures also is problematic for comparing across providers, regions, time periods, etc., if the denominators are substantively different. Thus, such measures may be best suited for internal quality improvement purposes.

Recommendations Regarding Data Included in Measures

Three of the 13 previously proposed potential solutions to the low case-volume challenge pertain to the data used in measures. In each case, the kernel of the proposed solution is to utilize (or "pool") additional data so as to increase the number of individuals included in the measure.

Pool Data across Time

As evidenced through the 2015 environmental scan, in statistical analysis, a common approach for increasing sample size (and thus alleviating the low case-volume problem) is to pool data on the same populations or samples over time (e.g., using three years of data rather than just one year).

The TEP agreed that the positive impact of this approach on sample size, and hence, on the power and reliability of analysis, is well established. This approach likely would provide opportunity to reliably assess performance for various patient subgroups (because sample sizes would be adequate). This approach also should result in increased, meaningful participation of rural providers in accountability programs.

On the negative side, the TEP recognized that pooling data across time may mask temporal changes (either improvement or degradations) in performance. This can be especially problematic when temporal change (in either direction) is of special interest. With this approach, measure results are less timely, and therefore less useful, for decision making by various stakeholders. This lack of timeliness may be particularly problematic if the goal is to include measures using this approach in programs that are specifically designed to incentivize performance improvements. In addition, different choices for the measurement time frame can lead to different results, particularly if the measure focus has been specifically targeted for improvement by certain providers. TEP members advised providing documentation to explain how to interpret results in general and when interpreting changes over time if pooling periods overlap. They also advised use of an adaptive and data-dependent approach for choosing the time frame for pooling when implementing this approach. The TEP expanded on this advice in their recommendations, as described in a subsequent section of this report that discusses the concept of "borrowing strength."

Aggregate Data from Multiple Providers

Aggregating over time, as described above, can be considered "vertical integration." An alternative to this approach is "horizontal integration." In such an approach, data from multiple providers (most likely those within the same regions or networks) are aggregated for the purposes of performance measurement. This approach is a potential solution to the low case-volume challenge that was identified in the 2015 environmental scan but was not discussed as such by the 2015 Rural Health Committee.

The TEP agreed that this type of aggregation also can effectively increase sample size, thus increasing statistical power, reliability, and the feasibility of stratified analysis. Provider aggregation would reduce the number of providers for whom results are displayed in public reporting programs, a result that might be appealing for some. Aggregation across provider groups also may result in increased heterogeneity of patient populations, which might facilitate subgroup analysis.

Similar to vertical integration, horizontal integration can also mask variation in performance across units that are pooled. Moreover, determining which providers to combine likely would be both subjective and arbitrary, and

different pooling schemes may lead to different analysis results and hence different decision making by users of the measures. As with pooling across time, results of measures that pool information across providers may be less useful for guiding quality improvement activities or other types of decision making, and they may have the unintended consequence of incentivizing individual providers to not pursue improvement efforts. Thus, as stated previously, use of this option would have to align with the specific goals of the measurement program.

Once again, the TEP offered an expansion of this approach in its recommendations, as described in a subsequent section of this report that discusses the concept of "borrowing strength."

Combine Inpatient and Outpatient Data for Similar Measures

Another approach for addressing the low case-volume challenge that emerged from the 2015 environmental scan is to combine inpatient and outpatient data for similar measures.

The TEP agreed that when inpatient or outpatient data are individually insufficient for reliable measurement due to low case-volume, pooling inpatient and outpatient data for comparable measures can be a sensible approach for increasing the effective sample size and would be technically feasible for various topics. This approach might be preferable for certain types of measures or topics, including resource use. In addition to increasing reliability and power, another potential gain from this approach is representativeness, in that observations from a wider spectrum of patients could be included. Moreover, both inpatient and outpatient data may be required to adequately assess certain conditions or treatments (e.g., surgeries performed in both the inpatient and outpatient settings for the treatment of breast cancer). TEP members acknowledged that attribution may be an issue when pooling data across inpatient and outpatient settings. However, they agreed that, for many measures, the accountable entity likely would be the provider involved in the index encounter.

On the negative side, the TEP noted that the imbalance between the numbers of inpatient and outpatient treatments (e.g., administration of chemotherapy) may cause one type of observation to dominate the analysis (this would vary by condition or topic area). TEP members also hypothesized that effects of inpatient-outpatient heterogeneity (e.g., in data, processes, patient mix, treatment options, etc.) may be undesirably or inappropriately eliminated, at least in some cases. Finally, decision making based on performance measure results may be more relevant to one care setting than the other; therefore, pooling data from both settings may hamper or otherwise invalidate the decision-making process.

Recommendations Regarding the Mechanics of the Measure Calculation

A final proposed solution for the low case-volume challenge that pertains to measure specifications is to use sophisticated statistical approaches (such as hierarchical modeling) when calculating healthcare performance measures. This approach emerged from the 2015 environmental scan but was not discussed by the 2015 Rural Health Committee. The legitimacy of the hierarchical modeling and related approaches, specifically, was established in the 2010 white paper[12] commissioned by the Committee of Presidents of Statistical Societies (COPSS), an effort funded by CMS.

The TEP agreed that there are more sophisticated (and potentially more powerful) statistical approaches that can be used to address the low case-volume problem than those presented above, with the hierarchical modeling approach[e] as a representative example. This methodology accounts for the inherent nested structure of the data (e.g., patients within hospitals) and supports both risk adjustment and stabilizing

e Other labels associated with some version or aspect of this approach include "multilevel modeling," "random effects modeling," "shrinkage," "reliability adjustment," and "borrowing strength."

provider-specific performance estimates by shrinking[f] estimates toward an appropriate target (which is usually a type of average). TEP members noted, however, that this target must be meaningful. The standard shrinkage target is the national mean or median, but other targets may be more appropriate, particularly for measures used to assess the performance of rural providers. For example, some may consider a factor that is related to procedure volume to be an appropriate shrinkage target. Regardless, measure developers should carefully consider use of provider attributes in producing shrinkage targets. The TEP expanded on these ideas in their recommendations, as described in a later section of this report that discusses the concept of "borrowing strength."

Potential Solutions that Address Reporting of Measure Results

Two of the 13 previously proposed solutions to the low case-volume challenge address how measure results are reported. One potential solution suggests options for reflecting the uncertainty of measure results, and the other recommends stratification according to provider type.

Present Confidence Intervals, Numerator Counts, and Denominator Counts

In statistical analysis, inference, as represented by confidence intervals, often accompanies point estimation. The TEP agreed that an indicator of uncertainty should be included when performance measure results are reported. Under the specific context of low case-volume, confidence intervals can be used to reflect the high level of uncertainty. Reporting additional information, such as numerator and denominator counts, if relevant, also can serve as a way to reflect uncertainty in measurement (i.e., very low numbers typically indicate lack of precision).

However, TEP members also noted that confidence intervals do not necessarily provide information that can assist decision making. In addition, they cautioned that confidence interval information can be misused if null-hypothesis significance tests are performed,[13] as these can be misleading if a measure is calculated using shrinkage techniques. The TEP offered an alternative recommendation regarding the presentation of uncertainty information, as described in a subsequent section of this report that discusses exceedance probabilities.

Stratify Providers so Performance Results are Compared Only among Similar Groups

The TEP noted that stratification is a common and highly useful way to eliminate variation, to potentially account for unobserved patient heterogeneity, and promote fairness in comparisons of providers when one compares "like to like." Members also remarked on the availability of statistical tools that facilitate stratified analysis.

However, the TEP also cautioned that stratification does not increase effective sample size and therefore does not actually address the low case-volume problem. Moreover, when multiple possibilities are present, choosing the proper stratification variable(s) can be challenging, and there is also risk of over-stratification. Perhaps most importantly, if rural providers are compared only to other rural providers, then it is not possible to determine whether rural providers are systematically providing higher- or lower-quality care than their nonrural counterparts.

f In small area estimation using multilevel models, part of the model structure is a regression model that predicts quality measure results from covariates representing structural characteristics of the unit being assessed (e.g., a rural hospital). Estimates of these models combine the prediction based on such characteristics with directly observed data on quality measures; the estimates that combine these two sources of information are more accurate than either source alone. Because the direct estimates, which are inaccurate and noisy (i.e., with substantial measurement error), are pulled in toward the relatively stable regression predictions, these are sometimes called "shrinkage estimates" and the structurally-based regression model predictions are called "shrinkage targets." In somewhat more technical terms, when performance estimates for a group of providers are based on a limited amount of data (known in statistics as small area estimation), the estimates will tend to be over-dispersed (i.e., be more spread out than the true long-term rates), because random measurement error is added to the true variation among providers. In general, a "stabilized" provider-specific performance estimate is a weighted average of the provider-specific estimate (i.e., the provider's risk-adjusted rate) and the estimate for a group of peers (e.g., providers with similar structural characteristics or attributes, or persistent past performance on related outcome and process measures—the "shrinkage target"). The weight that is applied is a "reliability adjustment" with a value from 0 to 1 (less than 1, hence the term "shrinkage" (signifying reversal of over-dispersion). This weight is generally calculated as the ratio of signal (true variation of provider rates) to total variation (the signal plus "noise" due to the limited amount of data). The more reliable the provider-specific estimate, the closer the weight is to 1; the less reliable the provider-specific estimate, the closer the weight is to 0.

TEP RECOMMENDATIONS

Using the above reflections on previously recommended potential solutions to the low case-volume challenge as a starting point, the TEP provided the following four recommendations to address this challenge.

Borrow Strength to the Extent Possible

Extending the ideas of pooling data across time, the TEP proposed a "partial pooling" approach that would increase reliability by combining data across time for some providers, to the extent that the data suggest differences across years are due to noise rather than signal. With this approach, data would be pooled over a longer period of time for providers with noisier data (i.e., those with more measurement error, such as low case-volume rural providers).

In this context, "partial pooling" is a synonym for "borrowing strength." Borrowing strength across time, for example, would strike a data-driven compromise between no pooling at all (e.g., the measure time frame is limited to a one-year lookback period for all providers) and complete pooling over an arbitrarily determined time frame (e.g., a three-year lookback period for all providers in order to ensure adequate sample size for all providers). This could be accomplished by incorporating data from previous years, but down-weighting those data relative to data from the current year, depending on how much noise there is in the data overall.

Ideally, the algorithm that determines the down-weighting would be consistent across providers, even though the weights themselves might vary across providers (i.e., some providers with sufficient volume or precision would have small or zero weights for earlier years, while providers with low volume and/or low precision would have higher weights for earlier years). This approach can be considered, conceptually, as "borrowing strength" for a particular provider from his or her past performance.

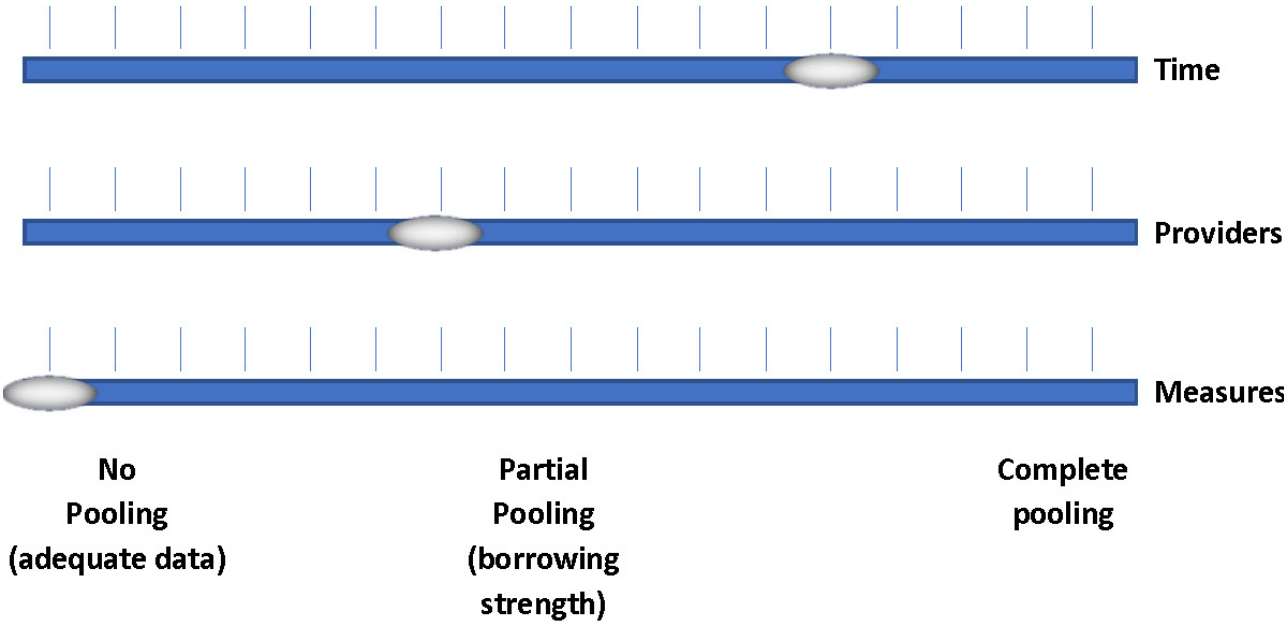
With such an approach, the available window of available data would be constant across providers (e.g., three years or five years, etc., whatever is practical in terms of implementation), although how much past data that would be used would vary across providers. Such an approach would be one way to implement the concept of "mastery," wherein one could set a reliability threshold and use as much past data as needed for a particular provider to reach that threshold.

The TEP then extended this concept by recommending pooling of other available information, such as by pooling data not only across years but also across similar providers (e.g., those with similar structural characteristics). Under such an approach, a provider's score would be informed both by his or her data from previous years and also by data from other providers with similar structural characteristics. The weights placed on these two sources of outside information would be determined entirely by the data. To be clear, this approach of borrowing strength across providers does not mean that data from other providers would be combined with that of a low case-volume rural provider, making it impossible to estimate the performance of that rural provider. Instead, information from similar providers would be used, again, as indicated through a data-driven algorithm, to improve the estimate for that rural provider, which otherwise would be inadequate due to low case-volume. As noted earlier, if borrowing strength is the analytic approach used to estimate performance, the user must understand that the resulting estimate is not an ideal substitute for what would be computed if data were unavailable and that care must be taken when making comparisons and interpretations based on these estimates. TEP members noted that this approach could be expanded even further by borrowing strength from other measures that are related to a particular measure, or by borrowing strength from other types of information such as data about procedures or conditions.

Figure 1 uses slider bars to illustrate the spectrum of approaches from no pooling of data, to partial pooling (i.e., borrowing strength), to complete pooling. The different bars signify various options

for pooling (e.g., across time, across providers, or across other measures). For any particular measure, any combination of borrowing strength might be appropriate (i.e., one may want to pool across time and providers but not across other measures), and there may be other dimensions in which pooling might be appropriate (i.e., there may be additional slider bars). When a rural provider has adequate data to estimate a particular performance measure, there would be no need for pooling (the slider bars would be positioned all the way to the left). But when there is a lack of data, partial pooling could be considered, with the amount of pooling (i.e., the optimal position on the slide bar) determined by the data, as described above. Complete pooling is an approach that might be considered depending on the purpose of measurement. As described earlier, complete pooling across time might mean using three years of data rather than just one year for all providers being measured. This approach might be appropriate if temporal changes in medical science, policy, etc., are minimal and assessment of improvement over shorter time frames is not a priority. As another example, pooling data across similar providers may be appropriate if the purpose is to assess some facet of overall quality of care, but not hold an individual provider accountable for that quality. Thus, the availability of data and the purpose of measurement determine the choice of no pooling versus complete pooling versus partial pooling (and the extent thereof).

FIGURE 1. ILLUSTRATION OF THE POOLING SPECTRUM



The TEP agreed that both outcome and process measures could benefit from this approach of borrowing strength but noted that that it would not be applicable to structure measures.[g] In addition, TEP members agreed that borrowing strength as a way of addressing the low case-volume problem is particularly important for "lower" levels of analysis (e.g., analysis at the individual clinician level). This is because the sample sizes inherent in clinician-level measures typically are substantially smaller than those in measures that assess "higher" levels (e.g., hospitals or health plans). Finally, the TEP noted that this approach of borrowing strength does not presume use of any particular data source or type of data (e.g., administrative claims, data from electronic health records, etc.).

One difficulty with the recommendation to borrow strength to the extent possible is that relevant or useful information may not be widely known. For example, some information may be known only to providers (e.g., patient-specific data or the provider's current performance results), while other information may be available only to implementers or sponsors of the measure (e.g., weights, model coefficients, the shrinkage target, the signal variance, etc.). However, TEP members believe this limitation is not insurmountable (e.g., a tool could be developed that provides some data and invites input of other data).

Another limitation of this approach is its complexity, making it difficult to understand without advanced statistical training. However, this potential limitation can be ameliorated by publishing the weights and shrinkage targets, so that individual providers could calculate their own performance scores.

TEP members also acknowledged the potential for temporal changes that should be considered when borrowing strength over time (e.g., changes in healthcare policy, methods of data collection, or in medical practice due to new guidelines, drugs, etc.). Again, they reiterated the previously-stated principle regarding the use and interpretation of results using this approach.

The TEP considered shrinkage estimation a form of "borrowing strength." TEP members stated a preference for using those indicators of structure with a strong link to the outcomes being assessed in a measure in defining shrinkage targets (e.g., having a catheterization lab when assessing AMI outcomes). They agreed that in some cases, using volume as a variable in defining the shrinkage target may be appropriate, but it may be overused simply because volume data are straightforward to obtain. They emphasized that volume is only one of many possible structural predictors, and that it may not be the best, particularly because the direction of causality is not always clear. For example, it may not be clear whether high case-volume results in high quality because "practice makes perfect" or whether providing high-quality care leads to high case-volume. Use of volume as a shrinkage target may be appropriate for the first scenario but not for the second. Thus, selection of appropriate structural predictors requires thoughtfulness and caution.

Finally, the TEP emphasized the need to consider the suitability of the particular approach in light of the use case for a particular measure. For example, if data are pooled across providers, the resultant measure results may be less useful for internal quality improvement efforts for individual providers. Similarly, if a program is designed to reward improvement between two time periods, it may not be appropriate to include measures that are estimated by pooling across time. TEP members suggested that program designers inform measure developers of how the programs will work, so that development aligns with the program design. Similarly, measure developers can clarify to program designers how their approach may or may not work for a particular program.

g Borrowing strength would not be applicable to structure measures because a particular provider would have the same value across all patients (e.g., a hospital's participation in a registry does not vary from patient to patient; a physician practice either does or does not offer a patient portal; etc.).

Recognize the Need for Robust Statistical Expertise and Computational Power

The TEP noted that the key challenge with the recommendation to borrow strength for low case-volume providers to the extent possible is in actually implementing the approach. TEP members agreed that implementation of this recommendation will require the professional expertise of PhD-level statisticians, not only to develop the statistical models needed to borrow strength, but also to write the necessary programming code to implement measures that include such models.

Implementing such measures also will require robust computational resources (i.e., computers with sufficient power to store, manage, and compute statistical models for very large datasets).

While acknowledging these substantial resource requirements, TEP members agreed that these should not deter the development of measures using the approach of borrowing strength. They also recommended initial pursuit of the most robust measures (i.e., those that maximize the amount of borrowed strength), even if a lack of statistical or computational resources ultimately necessitates a less vigorous approach.

Report Exceedance Probabilities

As noted earlier, use of confidence intervals to reflect uncertainty in measurement carries the risk of misuse if null-hypothesis significance tests are performed. TEP members suggested that use of exceedance probabilities could serve as an alternative way to reflect the uncertainty of measure results, without such errors. An example of an exceedance probability statement is the following: *We can be 84 percent sure that hospital A is performing above the mean on this particular measure*. The TEP noted that the 2012 COPSS white paper[12] recommended using exceedance probabilities when reporting performance scores and referenced more recent work by Shwartz, et al. (2014)[14] that demonstrates the utility of this approach for provider profiling.

TEP members articulated three advantages of exceedance probabilities. First, they reflect both the point estimate and its related uncertainty in one summary value. This may be particularly helpful in the context of performance measurement for rural providers (i.e., when uncertainty is high due to low case-volume), particularly if results of rural versus nonrural providers are being compared. Second, they summarize the posterior distribution (i.e., a Bayesian point estimate that has been shrunken or otherwise incorporates external information). Third, they are easily interpretable, particularly for consumers using measurement results to inform decision making.

The TEP also noted that the recommendation to use exceedance probabilities presupposes a view of performance that is more continuous than binary in nature. Thus, if the goal of measurement is to differentiate extremely good performance from extremely poor performance, exceedance probabilities might be of less interest. In contrast, if the goal of measurement is to help consumers (or others) maximize their chances of choosing a provider that would be most likely to provide a good outcome, then use of exceedance probabilities would be an effective way to present that information. Ultimately, TEP members agreed that the most effective choice for reporting provider performance hinges on the intended use of the measure from a policy perspective, as well as from the perspective of an individual user. Yet they also indicated that use of exceedance probabilities could serve as a best-practice reporting approach that could foster consistency across quality programs. Finally, because reporting exceedance probabilities along with performance measure results is still uncommon, the TEP also recommended that their use be paired with both education and field testing to ensure that healthcare consumers know how to interpret performance results.

Actively Anticipate Potential for Unintended Consequences

TEP members also noted the potential for performance measurement to drive decisions that can ultimately lead to unintended negative consequences. One example would be using a measure in an incentive program without realizing that it does not work well for rural providers. In this case, the unintended negative consequence might be misappropriation of incentive payments as a result of over- or under-estimation of the results for rural providers or encouraging activities that are counter-productive in rural environments (e.g., assessing timeliness of care but using those results to drive productivity efforts). Another type of unintended consequence might be using measurement results to drive large-scale policy decisions (e.g., regionalization of care) that affect rural residents and providers, but without proper consideration of potential downstream effects (e.g., decreased access to care for rural patients). In pointing out the potential for unintended consequences, the TEP emphasized that addressing the low case-volume problem is not simply a technical issue, but instead requires vigilance and a willingness to change course if needed. The TEP also agreed that formal feedback loops should be established to facilitate this vigilance.

RECOMMENDATIONS FOR FUTURE ACTIVITIES

In addition to the four key recommendations described above, TEP members also suggested several ideas for future work—potentially funded by CMS, either alone or in partnership with other entities—that could further address the low case-volume challenge for rural providers. Because these ideas have either a research or policy focus (or sometimes both), and therefore address different problems and potentially different audiences, the TEP did not try to prioritize them.

Research Activities

- Apply the recommendation of borrowing strength to the extent possible in a simulation study. In such a study, investigators could generate a simulated dataset based on the known true quality of a provider, then use various methods to estimate that known quality and see which method produces the best estimates (i.e., those closest to the known true quality). Investigation via a simulation study would foster both model development and statistical coding. One example for a simulation study would be to explore how measures using the borrowing strength approach would work if program scoring is based on improvement across time.
- Implement a "challenge grant" by providing either real or simulated data of rural providers with low case-volume—again where the true quality of the providers is known—and ask volunteer researchers to apply various methods to address the problem. The attractions of this approach include obtaining input from a variety of methodologists who most likely would use of a wide variety of methods.
- Explore which structural characteristics might be appropriate in defining shrinkage targets for performance measurement of rural providers. Such a project could include a literature review, followed by input from a TEP to discuss the literature and offer additional suggestions based on their own research or experiences. An empirical arm also could be included, perhaps modeling the most promising options via a simulation study.

- Bring together experts from other disciplines (such as education), who also must contend with the small denominator problem, in order to share best practices for measurement and reporting
- Explore nonparametric alternatives when developing measures for rural providers. The TEP recognized that parametric approaches predominate in healthcare performance measures but agreed that nonparametric alternatives should also be explored. The strength of such approaches is that they rely on fewer or less stringent statistical model assumptions, although they can be statistically less powerful than their parametric counterparts, particularly in relatively data-poor rural environments. The TEP pointed to the nonparametric approach of Bayesian Additive Regression Trees (BART).[15,16] This approach was recently proposed[17] as a way to assess the performance of teachers in the education setting (which is, statistically, very similar to healthcare performance measurement). This approach would allow measure developers to relax assumptions such as additivity and linearity that are used in more typical (parametric) approaches and possibly obtain better risk adjustment.
- Determine whether, and if so, how, to consider the small numerator problem, particularly from the rural perspective. The small numerator problem, which was considered out of scope by the TEP for this project, occurs when few individuals meet the measure numerator. Some measures should have small numerators (e.g., never events like wrong-site surgeries or rare events such as low-prevalence procedures), and there may therefore be implications regarding reliability, and, in turn, their use in accountability applications. But for rural providers, the small-numerator problem can also occur due to sparse population size. In such cases, rural providers' results for certain measures may be suppressed because of privacy concerns related to the low numerator counts.

Policy-Related Activities

- Explore the policy rationale for various approaches to measurement in rural areas, particularly considering quality improvement and access rather than competition.
- Explore the implications of lack of service delivery (e.g., obstetrical services; mental health services) in rural areas on performance measurement, particularly in the context of actual or theoretical pay-for-performance program structures.
- Revisit the core set of rural-relevant measures identified in 2018 by the MAP Rural Health Workgroup on an ongoing basis to ensure that rural residents and providers find these measures meaningful. Stakeholders involved in this effort should recognize that the selection of measures based on resistance to low case-volume may become less important as measures that borrow strength for rural providers are developed.
- Continue to explore ways to ensure that rural providers can meaningfully participate in quality programs, both public and private. As part of this effort, measures should be examined for their suitability for use in rural areas, as well as their suitability for use in accountability programs for rural providers. One example might be to ensure that the recommended analytic and reporting approach be tested and evaluated in demonstration programs that target rural providers.

CONCLUSION

After considering previously proposed potential solutions identified via an environmental scan and NQF's 2015 Rural Health Committee, the MAP Rural Health TEP provided four recommendations to address the low-case volume challenge for performance measurement of rural providers. These include borrowing strength to the extent possible, using statistical expertise and experts with robust computational skills to implement the recommendation of borrowing strength, reporting exceedance probabilities along with measure results, and actively anticipating the potential for unintended consequences of measurement for rural residents and providers.

As part of its deliberations, the TEP discussed the types of entities that could implement its recommendations. Given the complexity of the recommended modeling approach of borrowing strength, TEP members agreed that a national development and implementation effort likely would be needed. CMS or other federal agencies (e.g., AHRQ, HRSA), alone or in concert, would have the requisite capacity, contracting infrastructure, and data to spearhead such efforts. If, for example, CMS decides to carry out these recommendations, it could then decide more specifically how to do so. This could include deciding whether funding a national research group for this work would be logical, as well as making other contracting decisions such as the scope of measurement and required expertise.

Although not novel, the TEP's recommendations advance the field of healthcare performance measurement by addressing low case-volume, which continues to be a significant challenge for rural providers. This effort by the TEP uses a "rural lens" to spotlight the low case-volume challenge and its implications. The recommendation to borrow strength to the extent possible promotes Bayesian modeling as a preferred methodology for addressing the low-case volume challenge. As part of this recommendation, the TEP has advocated for a broader conceptualization of shrinkage beyond that of shrinkage to a national mean. The TEP's recommendation to consider more rural-relevant shrinkage targets in performance measurement for rural providers is particularly significant, given that shrinkage to a national average inhibits meaningful participation in CMS quality improvement programs by low volume rural providers (thus denying them the chance to benefit from the incentives of these programs). Finally, by couching these recommendations in more intuitive (rather than statistical) language, they are accessible to a variety of audiences.

REFERENCES

1 National Quality Forum (NQF). *Performance Measurement for Rural Low-Volume Providers*. Washington, DC: NQF; 2015. http://www.qualityforum.org/Publications/2015/09/Rural_Health_Final_Report.aspx. Last accessed January 2019.

2 National Quality Forum (NQF). *A Core Set of Rural-Relevant Measures and Measuring and Improving Access to Care: 2018 Recommendations from the MAP Rural Health Workgroup*. Washington, DC: NQF; 2018. http://www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=88226. Last accessed January 2019.

3 Rao JNK, Molina I. *Small Area Estimation. Wiley Series in Survey Methodology.* 2nd ed. Hoboken, New Jersey: John Wiley & Sons, Inc.; 2015.

4 United States Census Bureau. 2010 Census Urban and Rural Classification and Urban Area Criteria. https://www.census.gov/geo/reference/ua/urban-rural-2010.html. Last accessed January 2019.

5 United States Department of Agriculture. Economic Research Service State Data website. https://data.ers.usda.gov/reports.aspx?ID=17854. Published 2018. Last accessed January 2019.

6 Matthews KA, Croft J, Liu Y, et al. Health-related behaviors by urban-rural county classification — United States, 2013. *MMWR Surveill Summ*. 2017;66(5):1-8.

7 Moy E, Garcia M, Bastian B, et al. Leading causes of death in nonmetropolitan and metropolitan areas — United States, 1999–2014. *MMWR Surveill Summ*. 2017;66(1):1-8.

8 Meit M, Knudson A, Gilbert T, et al. *The 2014 Update of the Rural-Urban Chartbook*. Bethesda, MD: Rural Health Reform Policy Research Center; 2014. https://ruralhealth.und.edu/projects/health-reform-policy-research-center/pdf/2014-rural-urban-chartbook-update.pdf.

9 Douthit N, Kiv S, Dwolatzky T, et al. Exposing some important barriers to health care access in the rural USA. *Public Health*. 2015;129(6):611-620.

10 Southwest Rural Health Research Center, Texas A&M University. Rural Healthy People 2020 website. https://srhrc.tamhsc.edu/rhp2020/index.html. Last accessed January 2019.

11 National Quality Forum (NQF). *Guidance for Measure Testing and Evaluating Scientific Acceptability of Measure Properties*. Washington, DC: NQF; 2011. http://www.qualityforum.org/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=70943. Last accessed January 2019.

12 Ash AS, Fienberg SF, Louis TA, et al. *Statistical Issues in Assessing Hospital Performance*. Baltimore, MD: Centers for Medicare and Medicaid Services (CMS); 2012. https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/Downloads/Statistical-Issues-in-Assessing-Hospital-Performance.pdf.

13 Gelman A, Carlin J. Beyond power calculations: assessing type S (sign) and type M (magnitude) errors. *Perspect Psychol Sci*. 2014;9(6):641-651.

14 Shwartz M, Peköz EA, Burgess JF, et al. A probability metric for identifying high-performing facilities: an application for pay-for-performance programs. *Med Care*. 2014;52(12):1030-1036.

15 Hill JL, McCulloch R. Bayesian nonparametric modeling for causal inference. *J Comput Graph Stat*. 2011;20(1):217-240.

16 Chipman HA, George EI, McCulloch RE. BART: Bayesian Additive Regression Trees. *The Annals of Applied Statistics*. 2010;4(1):266-298.

17 Schiltz F, Sestito P, Agasisti T, et al. The added value of more accurate predictions for school rankings. *Econ Educ Rev*. 2018;67:207-215.

APPENDIX A:
TEP Members and NQF Staff

TEP Members

Mariel Finucane, PhD
Senior Statistician, Mathematica Policy Research, Inc.
Cambridge, Massachusetts

Jeffrey Geppert, EdM
Senior Research Leader, Battelle Memorial Institute
Columbus, Ohio

Shuangge (Steven) Ma, PhD
Professor of Biostatistics, Yale University
New Haven, Connecticut

Jessica Schumacher, PhD
Director of Data Management and Analytics for the Surgical Collaborative of Wisconsin,
University of Wisconsin–Madison, Surgical Collaborative of Wisconsin
Madison, Wisconsin

Alan Zaslavsky, PhD
Professor, Harvard Medical School
Boston, Massachusetts

NQF Staff

Elisa Munthali, MPH
Senior Vice President, Quality Measurement

Karen Johnson, MS
Senior Director

Suzanne Theberge, MPH
Senior Project Manager

Ameera Chaudhry, MS
Project Analyst

APPENDIX B:
Project Goals and Timeline

Between September 2018 and March 2019, NQF convened a five-member Technical Expert Panel (TEP) to develop recommendations to address the low case-volume challenge faced by rural providers.

Technical Expert Panel

The TEP comprised statisticians and measure methodologists. The Panel members were recruited via a national 15-day call for nominations. Their expertise includes proficiency in Bayesian statistics, small area estimation, nonparametric statistics, and performance measure development, as well as hands-on experience in quality measure reporting for low-volume rural providers.

Technical Expert Panel Deliberations

In October and November 2018, the **MAP Rural Health TEP** convened for three, three-hour conference calls to discuss the challenge of low case-volume for performance measurement in rural settings, review previously identified solutions to this challenge, provide additional solutions and recommendations, and plan the content of this report. In addition to their participation on conference calls, TEP members provided written feedback to follow-up questions after the calls and contributed to the writing of this report. The report summarizes the implications of the low case-volume challenge; the strengths and weaknesses of previously identified solutions; the TEP's recommendations for addressing the low case-volume challenge; and proposed next steps.

A draft version of this report was posted for a public comment period on January 18, 2018. The TEP reconvened in February 2019 to discuss comments received on the draft report and finalize their recommendations. The final report was released in March 2019.

Timeline and Deliverables

Month	Event
September 2018	Call for Technical Expert Panel Nominations
October 2018	Finalize TEP Roster TEP Conference Call #1: Project introduction and review of CMS quality programs; review of previous low case-volume recommendations; start discussion of new recommendations
November 2018	TEP Conference Call #2: Discussion of pros and cons of previously recommended solutions; discussion of additional statistical methods TEP Conference Call #3: Finalize TEP recommendations; discussion of additional cautions and considerations; ideas for future research and consideration
January 2019	Deliverable: Draft Report Public Comment Period
February 2019	TEP Conference Call #4: Post-Comment Call
March 2019	Deliverable: Final Report

APPENDIX C:
Public Comments Received on the Draft Report

American Medical Association

Koryn Rubin

The American Medical Association (AMA) appreciates the opportunity to comment on the National Quality Forum's draft report on "Addressing the Low Case-Volume Challenge in Healthcare Performance Measurement of Rural Providers: Recommendations from the MAP Rural Health Technical Expert Panel".

The AMA believes that the previous identified solutions outlined by this Technical Expert Panel (TEP) provide a clear picture of the benefits and risks of the various approaches and we remain supportive of the recommendations. While we do not necessarily disagree with the recommendations regarding data, we caution that solutions such as pooling data over time or aggregating data across multiple clinicians may have the unintended consequence of misrepresenting of true performance and the quality of care provided by one clinician versus another. Determining whether one clinician is truly driving improvements or not based on the quality delivered by a group of potentially heterogeneous clinicians could negatively impact the actionability of measures and data at the point of care.

The AMA urges the TEP to further consider what the implications of the proposed approaches such as borrowing strength may be when implemented in a program or to others such as small practices? For example, what impact could borrowing strength have if applied to the Merit-based Incentive Payment System (MIPS)? Would the results and associated Quality or Cost Category scores become less meaningful since they are not representative of recent performance? How effectively would an approach such as borrowing strength work in a program that will eventually assign points based on improvement in performance between the previous and current reporting years, particularly if a clinician has a low case volume and higher weights using past performance years or across clinicians are required? The AMA recommends that this scenario such as scoring for improvement in the MIPS Quality or Cost Category be one of the future simulation studies. Because the low volume challenge faced by those providing care in the rural setting is often similar to those encountered by small practices, it would also be useful if the TEP could examine whether these same solutions could assist small practices in programs such as MIPS.

We are supportive of the recommendation to include exceedance probabilities as we agree that this information would be very useful in informing clinicians, patients, and end users in how confident they can be about the results.

The AMA believes that the Centers for Medicare and Medicaid Services (CMS) and others who may begin to implement these recommendations must balance whether the approach (e.g., borrowing strength) is feasible, particularly in programs such as MIPS. If it is expected that this type of information must occur at a national level and is less likely to be feasible at a regional or local level, then are the approaches so complex that the resulting measures and data may be meaningless at the point of care to drive improvements? There is the potential for unintended consequences to occur with approaches that require significant infrastructure and resources. For example, could it lead CMS and others to resort to data sources that are easier to access and manipulate (e.g., administrative claims) rather than the more robust clinical data we all desire? The AMA asks that the TEP thoughtfully consider the downstream implications of these types of approaches when implemented in real world programs while finalizing this report.

Columbia University

Paul Kurlansky

Report is generally excellent and addresses a problem which arises in other circumstances but has particular aspects relevant to the rural setting, to which the panel members were sensitive.

Regarding specific issues raised:

- regarding weighting across time, panel members suggested that since volume of different providers may vary over time, the actual time "needed" for a particular provider to acquire sufficient data to meet the volume requirements of the metric might vary from provider to provider. There is, however, an inherent danger in this approach which does not seem to have been considered by the panel, in that medical care changes over time. Let us assume that in order to reach the threshold of x cases, provider A needs 3 years of data but provider B needs only 2 years. However, one year ago a new drug or surgical procedure was introduced that might impact the parameter being measured. Assuming a constant volume per year for the purposes of this example, provider A will only have had the benefit of this therapeutic advance for one third of his/her patients, while provider B will have had it for half of his/her patients. Therefore, caution needs to be exercised in varying the timeframe between providers.
- regarding pooling among providers, this approach may be particularly appropriate if providers actually work together, such as in the same clinical practice, or, more loosely, for the same hospital
- regarding the need for PhD level expertise in analysis of results, this is likely to be an absolute requirement. The data collection and reporting burden would not change or increase, but the analytical expertise will need to be available and this needs to be a clear commitment on the part of CMS if they wish to serve the interests of quality improvement in the rural setting
- I strongly applaud the circumspection of the Panel in raising the issue of unintended consequences—this process needs to be iterative—such consequences may be completely obscure at the time of metric implementation and only apparent over time. Therefore there needs to be a feedback mechanism that can assess this issue on an ongoing basis and be prepared to modify or change course
- I really like the idea of challenge grants--the same information or even metric might be analyzed differently by different groups of intelligent people and this might be a way of uncovering options that were not otherwise apparent

Federation of American Hospitals

Claudia Salzberg

The Federation of American Hospitals (FAH) appreciates the opportunity to comment on the National Quality Forum's (NQF) "Addressing the Low Case-Volume Challenge in Healthcare Performance Measurement of Rural Providers: Recommendations from the MAP Rural Health Technical Expert Panel" draft report. FAH appreciates the NQF's focus on this important topic and provides comments to further strengthen the report.

FAH supports the previously identified solutions but believes that the potential for misrepresentations of the quality of care provided should not be minimized when data is aggregated across multiple providers. This aggregation will require determinations of key characteristics to ensure that this pooled data presents as homogeneous a picture of quality as feasible but the validity of those methods must be demonstrated. FAH also questions whether this approach of aggregation would negatively impact the ability of providers to use these data and performance scores for quality improvement. The ability to broaden participation must not compromise our ultimate goal of driving improvements in patient outcomes.

In addition, FAH believes that additional work through multiple simulation studies are required prior to CMS or others moving forward with implementation of some of the recommendations such as borrowing strength. While these approaches are appealing given the potential to broaden participation of those with low case volumes such as rural health providers, it is not yet clear what the implications of an approach such as borrowing strength would have in programs that intend to measure improvements in scores and not just achievement of certain benchmarks. Could borrowing strength lead to less meaningful results since they are not representative of recent past performance

if higher weights are assigned to past years' scores or the mix of providers that are pooled prove to be invalid?

FAH is intrigued by the potential to use exceedance probabilities to better inform providers, patients and others. We suggest that additional testing and education is included as a part of this recommendation to ensure that its use achieves the desired end result – better informed consumers and providers.

There must also be a balance to many of these approaches to ensure that they are feasible, both by CMS in their programs and others. The report specifically states that many of the proposed approaches will require implementation at the national level, which leads FAH to question whether some of the recommendations are so complex that they may be meaningless at the point of care to drive improvements. There is also the potential unintended consequence that implementers will resort to data sources that are easier to access and manipulate (e.g., administrative claims) rather than the more robust clinical data such as electronic health record data, particularly if large data sets are required. It would be useful if the TEP could propose possible solutions to minimize this potential risk. FAH appreciates the opportunity to comment and asks the TEP to carefully consider our concerns and suggested improvements.

Stratis Health on behalf of The Rural Policy Research Institute

Jennifer Lundblad on behalf of Keith Mueller

The RUPRI Health Panel strongly supports the NQF Measure Applications Partnership Rural Health Workgroup. Valid and reliable quality performance measures are critically important to rural health care organizations. Low-case volumes may be the most challenging rural quality measurement issue; i.e., rural health care organization quality measurement has been statistically challenged by low case-volumes, and rural health care organizations have been historically precluded from many health care reform demonstrations due to low case-volumes.

The RUPRI Health Panel supports the statistical models recommended by the NQF Technical Expert Panel to address low case-volumes. However, we wish to underscore the recommendation that CMS, HRSA, AHRQ, and other quality measurement organizations specifically incorporate the proposed models in quality measure development, interpretation, and application within public scoring and payment incentive systems. Furthermore, we recommend that CMMI and other health care demonstration organizations incorporate the proposed statistical models into health care demonstration design and participant eligibility.

ISBN 978-1-68248-111-0
© 2019 National Quality Forum

NATIONAL QUALITY FORUM
1030 15TH STREET, NW, SUITE 800
WASHINGTON, DC 20005

http://www.qualityforum.org