The Significant Lack of Alignment Across State and Regional Health Measure Sets

Health Care Performance Measurement Activity:
An Analysis of 48 State and Regional Measure Sets

Kate Bazinsky and Michael Bailit
Bailit Health Purchasing, LLC
September 10, 2013

The Significant Lack of Alignment across State and Regional Health Measure Sets

Health Care Performance Measurement Activity: An Analysis of 48 State and Regional Measure Sets

Introduction: Private and public purchasers are interested in changing the way that they pay for health benefits by moving from a volume-based fee-for-service system to a system that pays for value. They argue that this movement will improve quality and reduce costs. Fundamental to value-based payment systems are performance measures that can assess to what extent the providers are achieving these twin goals. Recognizing that implementing the right measures is critical to the success of a value-based payment system, the Buying Value initiative, a coalition of employers, business health organizations and union health funds, has been working with CMS and private payers, i.e., health plans, to create a recommended measure set for use in such efforts. In beginning its work, the group quickly realized that before making a new contribution to the measurement world, it needed to have a better understanding of the measurement landscape across the states. Therefore, it commissioned Bailit Health Purchasing, LLC (Bailit) to conduct an analysis of a broad array of state-organized measure sets. Bailit gathered 48 measure sets representing different program types, and designed for different purposes, across 25 different states and three regional collaboratives.

Methodology: In identifying the 48 measure sets for the analysis, Bailit used a convenience sampling approach. Bailit requested assistance from contacts in various states, collected sets from state websites and solicited measure set recommendations from the members of the Buying Value initiative. It is important to note that Bailit did not survey every state, nor did it capture all of the sets used by the studied states. In addition, insurer and provider-organized measure sets were not studied. However, three measure sets from regional collaboratives were included in the analysis. Bailit, after consultation with Buying Value, excluded hospital-focused measure sets and removed 53 hospital-focused measures from the sets used in the analysis.

The goal of the analysis was to provide basic summary information to describe the 48 measure sets and assess the extent of alignment across the measure sets. The analysis sought to answer the following questions:

1. Are the measures used primarily standard measures?
2. To what extent are measures NQF-endorsed?
3. What are the primary sources of the measures?
4. Into which domains do most of the measures fall?
5. To what extent do the measures cover all age ranges?
6. To what extent are measures shared?
7. What are the most frequently shared measures?

Key Findings:

1. **There are many state/regional performance measures for providers in use today.** Across the 48 measure sets, we identified a staggering 1367 in use. When we looked at the "distinct"[1] measures across these sets, removing all of the duplicates across and within measure sets, we identified 509 distinct measures. Using the National Quality Strategy (NQS) tagging taxonomy developed by NQF, we found that these measures were distributed relatively evenly across all domains, with a focus on the "Treatment and Secondary Prevention" measures and the "Health and Well-being" measures. We also found that most measures are created for adults, but there does not appear to be a deficiency in the number of measures that could be used for the pediatric population or specifically for the age 65+ population.

2. **There is little alignment across measures sets.** State and regional measure sets don't "share" very many measures, meaning that they have very few measures in common. Only 20% of all measures were used by more than one program. Additionally the programs do not share these "shared measures" very often. No measure was used by every program. Breast cancer screening is the most frequently used measure and it is used by only 63% of the programs.[2] Only 19 measures were used by at least 1/3 of the programs.

3. **Non-alignment persists despite preference for standard measures.** Although 59% of the measures come from standard sources--with 52% of all measures coming from HEDIS—the programs are selecting different subsets of these standard measures for use.[3]

4. **Most programs modify a portion of their measures, which also contributes to a lack of alignment.** Even when the programs select the same measures, the programs often modify the traditional specifications for these standard measures. 83% of the measure sets contained at least one modified measure. 23% of the identifiable standardized measures were modified.[4] Two of the programs modified every single measure and six of the programs modified at least 50% of their measures.

---

[1] If a measure showed up in multiple measure sets, it was counted once (e.g., breast cancer screening was counted 30 times in the total measures chart since it appeared in 30 different measure sets, but was counted once as a "distinct" measure). If a program used a measure multiple times ("variations on a theme"), it was only counted once (e.g., the Massachusetts PCMH Initiative used three different versions of the tobacco screening measure; it is counted only once as a distinct measure).

[2] Ironically, this measure is no longer NQF-endorsed due to changes in clinical guidelines put forth by national organizations.

[3] Please note that some of the measures included as "standard measures" have been modified by the programs.

[4] In this analysis, if Bailit did not have access to the specifications, but the measure appeared to be standardized through combination of steward and title or NQF#, it was considered to be a standard

5. **With few exceptions, regardless of how we analyzed the data, the programs' measures were not aligned.** Even though the measures used were selected from the same domains, the programs did not select the same measures from within each domain. This suggests that simply specifying the domains from which programs should select measures will not facilitate measure set alignment. Additionally, while one might hypothesize that programs designed for the same type and/or purpose would have more similarities than the full array of studied measure sets, this is not the case. Bailit reviewed four different types of programs (13 patient-centered medical home (PCMH) programs, six Medicaid MCO programs, six "other provider" programs,[5] and three regional collaborative programs) and found that only Medicaid MCO programs shared more, rather than fewer measures, sharing 62% of their measures. However, none of the other types of programs showed much alignment. The "other provider" programs shared the least number of measures, with only 12% shared. Additionally, while one might anticipate that programs developed for payment would be more standardized and be comprised of primarily NQF-endorsed measures, this is not the case. We also looked at the measure sets within two states: California (CA) and Massachusetts (MA) and found that CA has significantly more alignment across its measure sets when compared to MA and the total measures set. This alignment within CA may be due to our sample; three of the seven measure sets were developed by the same organization (Office of the Patient Advocate). However, anecdotally, we have been told that CA has also worked to align its measure sets. While MA has work underway to align its measure sets across the state though the Statewide Quality Committee, currently there is little alignment within the state.

6. **Many programs create their own measures.** 40% of the programs created at least one new measure for use, resulting in 198 homegrown measures. Of these 198 measures, there were 28 measures (14%) for which it was not readily apparent as to why the program created the measures, as these measures appeared to replicate standard measures. Perhaps the programs were unaware of the availability of the standard measures. 41% of the measures were specific to an aspect of a particular program. These measures primarily related to infrastructure, utilization, geographic access and oversight of a program. Since they are specific to the management or structure of a particular program, they are unlikely to become standardized. Approximately 10% of the measures appear to have been designed to give providers additional flexibility and options with regard to the measurement tool or outcome. For example, the Texas program includes a quality of life measure, but allows the provider to select a validated tool to offer its providers flexibility regarding which tool they use. Finally, approximately 35% of the homegrown measures seem to fill a perceived measurement

---

measure. This approach is likely to underestimate the extent of modification and suggests that there is likely to be more modification than represented by this analysis.

[5]The "other provider" category denotes programs that are focused on either paying or reporting performance at the provider level, but are not used for ACO, PCMH or Health Home programs.

gap. These measures focused on the areas of care management and coordination, patient self-management and cost.

7. **Most homegrown measures are not innovative.** While we found that most of the innovation in the measure sets came from the homegrown measures, most of the homegrown measures were not particularly innovative. For this analysis, Bailit defined **"innovative"** to describe measures that are not NQF-endorsed and that address an important health care concern that is not addressed in most measure sets (e.g., care management and coordination, cost, end-of-life care, patient self-management, social determinants of health) or address an issue or condition for which few measures are commonly employed (e.g., dementia, dental care, depression and other mental health conditions, maternal health, pain, quality of life, and substance abuse). Innovation is not widespread across measure sets. Only 38% of the programs included innovative measures and most programs only included one or two innovative measures. Only two programs did a significant amount of innovating in measurement:[6] the Massachusetts PCMH Initiative (17 measures) and the Texas Delivery System Reform Incentive Program (17 measures). Bailit also reviewed two additional measure sets from regional collaboratives that were not included in the core analysis to identify whether they would offer more innovation (Minnesota AF4Q and Oregon AF4Q). Oregon did not include any and Minnesota included four innovative measures.

8. **There appears to be a need for new standardized measures in the areas of self-management, cost, and care management and coordination.** Programs tended to focus their innovation efforts in these areas, suggesting a need for new standard measures in these arenas.

Conclusion: The expansion in the number of standardized measures affords state entities and regional collaboratives many more options than were available two decades ago. Programs tend to use these diverse measures to create their sets independently without an eye towards alignment, focusing rather on their particular local, programmatic needs as well as the desires of various constituencies, such as medical specialties. Even those who may seek alignment across measure sets will find few tools available to help facilitate this alignment. This lack of alignment is burdensome to providers who must report large numbers of measures to different programs and meet related accountability expectations and performance incentives. Mixed messages about quality priorities from these various measure sets results in "measures chaos" and makes it difficult for providers to focus their quality improvement efforts. It is also frustrating to purchasers and payers who seek to align incentives and market signals across disparate programs and markets.

[6] While we were not able to review the specifications for the Texas measures, some of these innovative measures appear to be measure concepts that do not yet have specifications, rather than actual measures that are ready to be implemented.

We anticipate that as states and health systems become more sophisticated in their use of electronic health records and health information exchanges, there will be more opportunities to easily collect outcome-focused, clinical data-based measures and thus increase use of those types of measures over the traditional claims-based measures. Combining this shifting landscape with the national movement to increase the number of providers that are paid for value rather than volume suggests that the proliferation of new measures and new measure sets is only in its infancy. In the absence of a fundamental shift in the way in which new measure sets are created, we should prepare to see the problem of unaligned measures grow exponentially.

Recommendations: In order to address the problem of measures non-alignment, we recommend the following strategies:

1. **Launch a campaign to raise awareness about the current lack of alignment across measure sets, help states and regions interested in developing measure set understand why lack of alignment is problematic and establish the need for a national measures framework.** In the absence of such an initiative, states and regions interested in creating measure sets have worked and are likely to continue working independently without an eye towards alignment at a state, regional or market level.

2. **Communicate with measure stewards to indicate to them when their measures have been frequently modified**, in particular in the cases in which additional detail has been added, removed or changed (i.e., not when the program just chose to report one of the two rates included in the measure). We recommend sharing with the measure stewards the specific types of modifications that have been made.

3. **Develop an interactive database of recommended measures to help establish a national measures framework.** While the criteria for inclusion in this interactive -- preferable online — tool would have to be more clearly defined, we recommend that it consist primarily of robust standardized measures that are used most frequently for each population and domain. We also recommend identifying measures for the areas in which there are currently few, if any, standardized measures (e.g., patient self-management, care management, cost, etc.). In order to be a success, this resource should be marketed to public and private sector organizations involved directly or indirectly in measurement set development and updated regularly (i.e., at least on an annual basis).

4. **Provide technical assistance to states to help them select high-quality measures that both meet their needs and encourage alignment across programs in their region and market.** This assistance could include:
    a. a measures hotline that states, regional collaboratives and engaged stakeholders could call to ask questions about:
        i. the use of the interactive measures tool;
        ii. help selecting appropriate measures; and/or
    a. learning collaboratives, blogs, online question boards and/or listservs dedicated to individuals working with measure sets.
    b. the creation of benchmarking resources for the recommended measures selected for inclusion in the interactive measures tool. Programs seeking to implement

quality measures to evaluate performance often struggle to set appropriate targets in the absence of benchmark data. While NCQA provides this information for its HEDIS measures through its Quality Compass® tool, benchmarks are not available for most other measures. Providing this information for those measures which currently lack benchmarks will provide programs with an important incentive to choose Buying Value measures over other measures.

5. **Acknowledge the areas where measure alignment is potentially not feasible or desirable.** It is important for the developers of measures sets to consider their populations of focus. For example, we would not recommend that a commercial measure set be identical to a Medicaid measures set that is designed to assess performance of long-term support services in a dually eligible (i.e., Medicare and Medicaid-eligible) population. Additionally, we anticipate the programs will continue to use program-specific measures, especially those that are administratively-focused (e.g., the rate of PCMH enrollment).