Key Design Considerations in Physician Value-based Purchasing

Cheryl L. Damberg
Senior Researcher, RAND Corporation

Physician Value-Based Payment Modifier
Special National Provider Call:
Experience from Private Sector Pay-for-Performance Programs

March 14, 2012

The views expressed in this presentation are those of the speaker and do not represent the views of CMS
Overview of Presentation

- Desired features of performance-based accountability systems: lessons from various business sectors

- Design elements in Value-Based Purchasing (VBP) programs
  - Incentives
    - Size and structure of incentive payment
    - Lessons from the behavioral economics literature on structuring the incentive payments
  - Measures and measure properties
    - Reliability and risk of misclassification
    - Validity
    - Composites
  - Complexity
  - Distribution of awards
Performance-based Accountability Systems (PBAS)*

• PBAS—of which Pay-for-Performance (P4P) and value-based purchasing are examples—can be an effective strategy for improving the delivery of services to the public

• Optimum circumstances include having the following:
  — A goal that is widely shared
  — Measures that are unambiguous and easy to observe
  — Incentives that are meaningful to those being incentivized
  — Few competing interests or requirements
  — Adequate resources to design, implement, and operate the PBAS

• Note: these conditions are rarely fully realized

**Performance-based Accountability Systems (PBAS)**

- PBAS designers face three basic design issues:
  - Determining whose behavior they seek to change (i.e., identifying individuals or organizations to target)
  - Deciding on the type and size of incentives
    - Which incentive structures to use to affect the desired behaviors?
  - Measuring performance and linking these measures to the incentives they have chosen

- Context can have a large effect on the incentive structures that PBAS designers choose

- Unclear how well the magnitude of rewards is correlated with the benefits of the changes that the PBAS designers seek to induce or the effort that service providers must make to comply with these changes

The measures used to quantify performance can vary in many dimensions. A need to consider a number of competing factors when selecting and structuring measures:

- The feasibility, availability, and cost of measures
- The context within which a PBAS operates
- The alignment of measures with PBAS goals
- The degree of control of the monitored party
- Resistance to manipulation by the monitored service activity
- Understandability

The selection of performance measures ultimately requires some trade-offs among these factors

Measures*

• Measures dictate the things on which providers should focus and what they might choose to ignore or neglect

• Create “risk-adjusted” output measures that account for relevant social, physical, and demographic characteristics of the population served (a threat to validity of measure)

• Measure relative improvement rather than absolute performance
  
  — Avoid measures that focus on a single absolute threshold score
  
  — Low achievers with no realistic prospects for achieving the absolute threshold score will have no incentive to seek even modest improvements

  — High achievers will have no incentive to strive for further improvement

  — Alternatives include use of multi-threshold scores and measurement of year-over-year improvement

Structuring Incentives: Lessons from Behavioral Economics

• A series of small incentives is better than one large incentive
  — It may be more psychologically motivating to provide a physician with smaller and more frequent incentive payments than a larger single lump-sum incentive payment

• A series of tiered absolute thresholds is better than one absolute threshold
  — An individual’s motivation and effort when faced with a goal greatly depend on his or her baseline performance (“goal gradient”)
  — If baseline performance is far away from goal performance, the individual exerts little effort because the goal is viewed as not immediately attainable.
  — A greater behavioral response is likely when there are a series of quality performance thresholds to meet (e.g., increasing amounts of money for achieving 50%, 60%, 70%, 80%, and 90% performance thresholds) rather than one (e.g., a 75% performance threshold)

Structuring Incentives: Lessons from Behavioral Economics

• Reducing the lag times between care and receipt of incentives increases the behavioral response
  — Money received right away is perceived as different in value than money to be received in the future, even the near future

• Although withholds have more of an effect than bonuses, one needs to be cognizant of the negative psychological response
  — Individuals are more sensitive to incentives when they perceive that they are losing something as opposed to gaining something
  — Incentive payments can be structured as a withhold (perceived loss) or as a bonus (perceived gain)

• Reducing the complexity of an incentive plan increases the behavioral response
  — Individuals often cannot process complex decisions that are tied to a financial incentive

• “Teaching to the test”
  — A physician’s “output” includes many different components (multi-tasking) such as managing a patient’s chronic illness, timely and efficiently diagnosing a patient’s new symptom, and counseling and advising on how to prevent illness
  — However, incentive programs often address only a narrow portion of a physician’s outputs or the processes that contribute to outputs
    • A large financial incentive based on a narrowly focused set of measures may lead to the unintended consequence of having a physician focus on items being measured and neglecting other important outputs that are not being measured
    • Mitigate by addressing an extensive array of a physicians’ output by applying a broad dashboard of performance measures

### Summary: Seven Design Features That Could Improve P4P Programs

<table>
<thead>
<tr>
<th>Commonly Used Design</th>
<th>Suggested Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incentive given as a lump sum</td>
<td>Divide the lump sum into a series of smaller incentive payments</td>
</tr>
<tr>
<td>Relative thresholds (e.g., top 25% of physicians)</td>
<td>Use tiered absolute thresholds (e.g., 25%, 50%, 75%, and 90%)</td>
</tr>
<tr>
<td>Long lag time between care and receipt of incentive</td>
<td>Shorten lag time to as short as possible</td>
</tr>
<tr>
<td>Use of withhold payments</td>
<td>Consider bonus payment or use of deposit contracts</td>
</tr>
<tr>
<td>Complex uncertain structure of program (e.g., shared savings program)</td>
<td>Simplify program so that uncertainty is minimized</td>
</tr>
<tr>
<td>Incentive often given as an increase in fee schedule reimbursement</td>
<td>Decouple incentive payment so that it is given separately; consider a lottery</td>
</tr>
<tr>
<td>Monetary incentives</td>
<td>Use in-kind incentives</td>
</tr>
</tbody>
</table>

Misclassification refers to reporting a provider’s performance in a way that does not reflect the provider’s true performance. For example, provider’s performance may be reported as being in category “1” when true performance is in category “2.”

Misclassification is related to:

- The reliability of a measure
  - Which depends on sample size (which can vary provider to provider)
  - Variation between providers (so population dependent)
- Number of cutpoints in the classification scheme
- How close the performance score is to the cutpoint

Between-Provider Performance Variation

Lower between-provider variation
(harder to tell who is best)

Higher between-provider variation
(easier to tell who is best)

= average performance for each provider
Different Levels of Measurement Error

(Uncertainty about the “true” average performance)

Higher measurement error (harder to tell who is best)

Lower measurement error (easier to tell who is best)

● = average performance for each provider

= range of uncertainty about “true” average performance

Measurement Issues: Using Composites to Help with Small Numbers

- Providers likely to have a small number of patients who qualify for any given measure
  - With the exception of cancer screenings, immunizations, and high prevalence chronic conditions
- Composite measures are one method to improve your ability to get a good signal on the estimate of a provider’s performance
  - Can potentially increase reliability*
- Higher reliability in a measure:
  - Means more signal, less noise
  - Reduces likelihood that you will classify provider in “wrong” category
- Per Adams**: “Reliability ASSUMES validity”

Sources: * David Reeves, PhD, Stephen M. Campbell, PhD, John Adams, PhD, Paul G. Shekelle, MD, PhD, Evan Kontopantelis, PhD, and Martin O. Roland, DM. Med Care 2007;45: 489–496
Composite Example: Diabetes Composite (% of PCPs >.70 Reliability)

- Composite: 74%
- EYE: 1%
- HBA: 27%
- LDL: 27%
- MPM: 43%

Reliability: > 70%
Desire to pay for both attainment and improvement

- Total cost of care attainment
- Consumer Price Index (CPI) trend
- Physician group’s year-over-year trend

3 metrics make it difficult for a physician group to relate performance to payout

- Reduces ‘line-of-sight’ and incentivizing power
- Providers need focused and clear targets to translate targets into internal behaviors

Simpler model can reward attainment and improvement

- Consider 2-Component model if both are focus areas

3rd reward component--POs trend over previous year below its historical trend-- allows high cost/high quality groups to “win” and delays ‘bending the cost curve’
State versus Regional Targets Reward Different Segments of the Cost Distribution
Summary: Key Design Considerations

• Are incentives sufficiently large to induce behavior changes?

• Is the design sufficiently simple so that providers have a line-of-sight from their behaviors to rewards?

• In VBP, overly narrow market definitions risk rewarding relatively high cost providers

• Will the program punish providers who serve more challenging populations, and how can those effects be moderated either at the front end (risk-adjustment) or at the back end (stratification)?

• Do providers know their relative cost positions (in comparison with each other)?

