

## Best Practices in Physician Incentives and Rewards Programs

---

In 2004, Bridges To Excellence (BTE) and The Leapfrog Group published a White Paper on Measuring Provider Efficiency<sup>1</sup>. As the document's Foreword states: "While the paper focuses on efficiency, all the contributors acknowledge that measuring efficiency should be done in conjunction with measuring effectiveness of care, so that consumers, purchasers and payers can better understand and identify the value of the services being delivered, and providers can better understand the steps they need to take to improve the value of services offered."

Since then, the field of efficiency measurement has evolved and many of the methodological shortcomings identified in the White Paper have been documented in other studies<sup>2</sup> and, to some extent, addressed by the main suppliers of software applications that enable the comparative measurement of physician efficiency. In addition, as pay-for-performance programs have entered the mainstream of physician payment programs used by payers in their network contracts, the scrutiny of the methods used to derive comparative performance scores has greatly increased. Two important reports<sup>3</sup> have emerged in the last two years that have influenced the provider and payer communities: A report commissioned by the Massachusetts Medical Society on comparative performance reports generated in that state by the state's employee benefit program, and a series of letters issued by the NY state Attorney General's office on health plan-generated comparative performance rankings which resulted in settlements by the plans accepting scrutiny over their methods.

As a result of all these reports and activities, payers and purchasers are left to wonder how incentive and rewards programs can be designed in a manner that achieves their primary goals of improving the effectiveness of care services delivered to patients and the efficiency with which those services are used. This document is intended to provide such guidance by pulling from best practices observed throughout the country as well as BTE's own experience in implementing incentives and rewards programs for over five years<sup>4</sup>. Overall, and quite simply, there are two essential ingredients to a successful incentive and reward program: (1) the amount of money pooled as an incentive has to be commensurate to the level of effort required to qualify for the incentive, and (2) the data on which the performance achievement is determined have to be credible to providers. We will address each of these sequentially.

---

<sup>1</sup> <http://www.bridgestoexcellence.org/whitepaper>

<sup>2</sup> [http://www.medpac.gov/publications/congressional\\_reports/Jun06\\_Ch01.pdf](http://www.medpac.gov/publications/congressional_reports/Jun06_Ch01.pdf), and [http://www.pbgh.org/news/pubs/documents/PBGHP3Report\\_09-01-05final.pdf](http://www.pbgh.org/news/pubs/documents/PBGHP3Report_09-01-05final.pdf)

<sup>3</sup> [http://www.massmed.org/GIC\\_Review](http://www.massmed.org/GIC_Review), and [http://www.oag.state.ny.us/press/2007/oct/oct29a\\_07.html](http://www.oag.state.ny.us/press/2007/oct/oct29a_07.html)

<sup>4</sup> Three important papers summarizing BTE's impact are in pre-publication review: (a) de Brantes et al, Value of Ambulatory Quality Care Measures: A Payer-purchaser's Perspective, (b) Rastogi et al, Rewarding High-quality Care: Could There be a Business Case, (c) Rosenthal et al, Bridges To Excellence: Recognizing High Quality Care

#### A. Money Matters: An analysis of critical mass

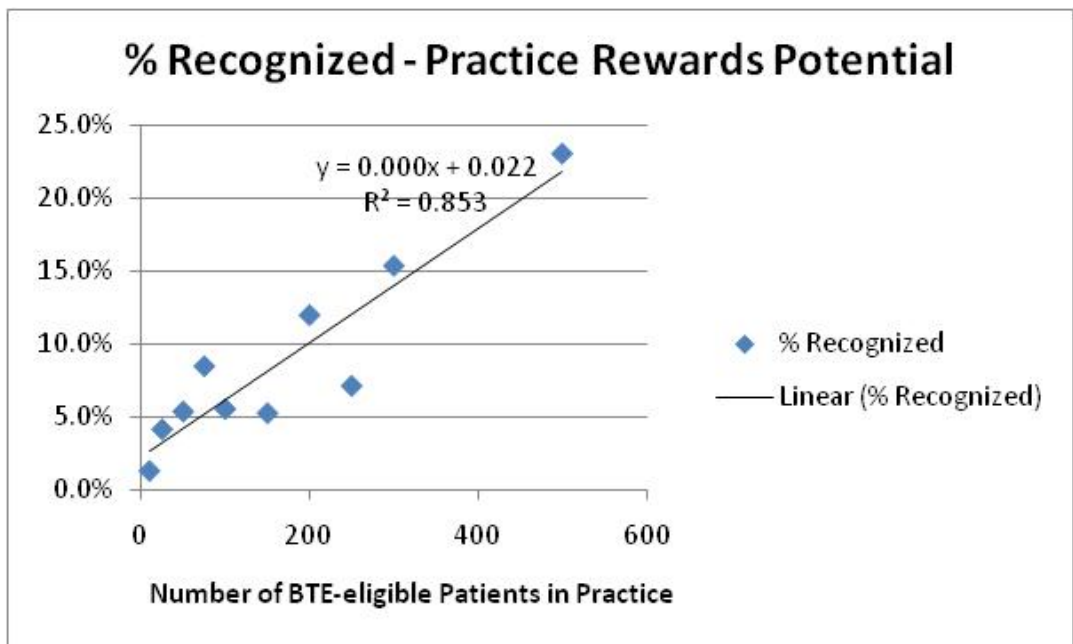
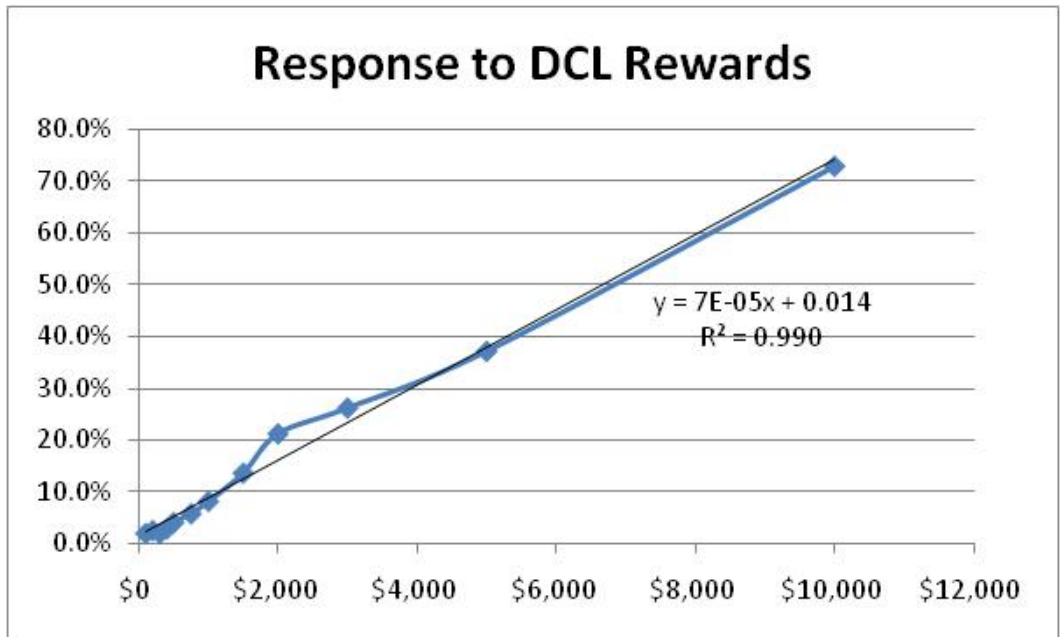
In December 2001, the now-defunct National Health Care Purchasing Institute published a report on the key criteria required by any incentive and reward program targeting physicians in order for that program to be successful<sup>5</sup>. Since then, many other reports<sup>6</sup>, including the reports from the Robert Wood Johnson Foundation's Rewarding Results program – in which BTE participated – and an Agency for Healthcare Research and Quality-funded report by Meredith Rosenthal and Adams Dudley, have simply reinforced this initial report. In addition, the example provided by the Integrated Healthcare Association in CA shows how the coordination of efforts by health plans in a state can impact provider behavior. Embedded in all these findings are some very practical criteria that are necessary to get physicians to respond to financial incentives:

- a. "The juice has to be worth the squeeze" – Physician practices, whether large or small, make investment decisions in a manner that is similar to the way most businesses and consumers make those decisions: they weigh the costs and benefits, and if the benefits are at least equal to the costs, then they are likely to make the investment. In BTE's overall program evaluation, we demonstrated that there was a statistically significant difference in the overall size of the incentive for practices that participated in the Physician Office Link (POL) than for practices that participated in the Diabetes Care Link (DCL). The difference stems from the size of the investment (costs) needed to qualify for the reward. For DCL, physicians have to target one class of patients and develop standard care protocols. For POL, physicians have to adopt a series of systems and processes all the way to and including an electronic medical record. As such, there is a significant difference in the cost to the practice for qualifying for DCL as opposed to POL. The two charts below illustrate the point: 20% of physicians will participate in DCL if the total rewards are \$2000 per year or greater, but it takes \$25,000 per year or more to get 20% of physicians to participate in POL (each patient is worth \$50 per year in rewards for POL). Why will some physicians decide to participate at lower incentive thresholds? Some because they've already achieved the required level of performance and the incremental cost of being recognized is small. Others because they are willing to take the risk that there will be more incentives of the same kind in the future. And still others because they belong to corporate structures that are willing to make a long term investment in the redesign of care. However, overall, those groups only represent 10% to 15% of the physicians in the community. The balance need incentives that will reasonably cover the costs of transformation. The amounts in our analyses (\$2,000 for single condition improvement, \$25,000 for total practice transformation) should be reasonable goalposts between which a plan should define its total incentive and reward program, both as a stand-alone and in conjunction with other P4P efforts in the community.

---

<sup>5</sup> <http://www.academyhealth.org/nhcpi/incentives.pdf>

<sup>6</sup> <http://www.ahrq.gov/qual/pay4per.htm>, and <http://www.bridgestoexcellence.org/Documents/BTE-Program-Evaluation-7-26-06.pdf>



- b. “The amount of the juice has to be predictable” – There are essentially two types of formulae for awarding incentives: tournament-style, or threshold-based. The former are similar to a race because you cannot predict in advance where you will end up, you only know once the race is over and the rankings created (e.g. you’re compared to peers). This is the method most commonly used to rank physicians on their cost-effectiveness. The latter is similar to most corporate compensation policies where specific goals are set in the beginning of the year and incentives are tied to each goal. This is the method that BTE has used from its inception and

recommends as a best practice for quality-based incentives. Our analysis clearly shows that true practice transformation requires a significant investment in time and money from the physicians, and, as a result, the benefits have to be quantifiable and predictable to get physicians to be willing to invest.

- c. “The more oranges there are, the more juice you get” – In many instances a single health plan does not have a significant enough market share of a physician’s panel to achieve the critical mass of dollars that will cause the physician to invest in care transformation. As a result, for that plan to have an effective incentive and reward program, it should use measures and performance assessment definitions that are the same as that of its competitors and other payers. While there is a risk of “free ride” by using standard performance assessments, that risk should be carefully weighed against the risk of poor care performance. The best practice in the industry is to use standard performance assessments and common metrics across a community<sup>7</sup>.

#### B. Data Credibility Matters: The tyranny of the “n” and how to attribute it.

The most ubiquitous, plentiful, and inexpensive data at hand to assess physician performance are claims and other administrative data. There, however, several limitations inherent to these data when measuring care effectiveness, and several pitfalls to avoid when measuring care efficiency.

- a. Best Practices in Measuring Care Effectiveness:
  - i. The importance of independence – one constant in an ever-changing health care system is the tension/conflict between providers and payers. It’s constant, because it’s inherent in the third-party payer system in the US. Plans negotiate contracts with providers that define the terms and price of the delivery of services. As a result, payers rarely trust providers to set rules and providers don’t trust plans in setting rules. Across the country, the most effective quality measurement programs are ones that are independent of the sole control of one or the other of these two parties. Independent third parties that assess physician performance using standard, nationally endorsed measures is the only way to avoid wasted years of arguing about the legitimacy of the assessment.
  - ii. The importance of sample sizes – Another advantage of independent third parties is that they might benefit from the aggregation of data across a physician’s entire patient panel, whether the data are administrative or medical record-based. Some of the reports listed above suggest that minimum sample sizes for performance assessment of physicians across multiple measures is 100 patients. In its most recent work in developing the comprehensive Care Practice Improvement Module (PIM)<sup>8</sup>, the

---

<sup>7</sup> <http://www.mnhealthcare.org/~main.cfm>

<sup>8</sup> <http://www.abim.org/pims/default.aspx>

American Board of Internal Medicine (ABIM) sampled 100 patients for each physician. For its recognition programs the National Committee for Quality Assurance (NCQA) uses a sample of 25 patients<sup>9</sup>. That's the same number used by the ABIM for its regular PIMs. In her article on BTE, Meredith Rosenthal observed that some physicians did not have an adequate sample size for some of the quality metrics that were being assessed through administrative data (even though claims from four plans were aggregated), while those same physicians did get quality scores through the medical record review process. The Best Practice in the industry suggests that for any single condition (e.g. diabetes, CVD, HTN) the minimum patient sample size should be 25, and that to measure a cross-section of the physician's panel on a number of conditions/diseases/care areas, the minimum sample should be 100.

- iii. The importance of outcomes – 150 years of industrial revolution, and 50 years of total quality management have taught all industries and services that outcomes are the only measure that matters to the buyer of the service. In health care, that certainly holds true for the patient. A research project commissioned by BTE and performed by Towers Perrin shows that there are a few handfuls of ambulatory care measures that have both clinical and economic value. The common element of these measures is that they are very good predictors of risk-reduction – if you reduce blood pressure in patients with hypertension, you will significantly reduce the risk of a heart attack and/or stroke. The dependence on administrative data has caused payers and providers to focus almost exclusively on measures that don't have a tight relationship to outcomes and, as a result, are poor predictors of risk reduction and cannot be associated with significant health care cost savings. Here again, communities that have been very successful in improving the quality of care of patients have focused almost exclusively on measures that relate to real patient outcomes. As such, health plan incentive and reward programs should give significant weight to outcome-based measures, including the patient's experience with care received.
- iv. The importance of choice – A network-wide incentive and reward program needs a mechanism and data to assess the performance of all physicians, not just those that volunteer. As such, and unless the plan is in a community that has a community-wide measurement organization that captures medical record and administrative data, the plan has to use administrative data as a common base in that performance assessment. However, the plan should build its assessment using the formula depicted below.

---

<sup>9</sup> <http://web.ncqa.org/tabid/74/Default.aspx>



## Health Plan Incentives and Rewards Program Formula

$$Q_p + \$_p = \text{Score}_p \rightarrow \text{Incentives}_p$$

Where the Quality score for each physician/practice =

$$f(\alpha Q1, \beta Q2, \gamma Q3, \delta Q4)$$

$$\alpha + \beta + \gamma + \delta = 100\%$$

Q1 = Claims/Admin data-based score

Q2 = Medical Record data-based score

Q3 = Patient Experience of Care-based score

Q4 = Hospital score

The Best Practice in determining the  $\alpha$ ,  $\beta$ , etc... weights of the Q1, Q2, etc... variables is to give a weight that is greater than 51% to  $\beta$  when Q2 is available. That's because Q2 has more outcomes information than Q1, and Q2 has additional attributes that are listed below. By creating a multi-variable care effectiveness score, the health plan is offering its network physicians a choice on the data source for that score. It's no longer a one size fits all. Instead, for physicians that can and want to provide medical record-based data, that option is both available and encouraged through the weighting system. This formula explicitly acknowledges that there are shortcomings to administrative data and encourages the provision of more reliable data.

- v. The importance of credibility – Physicians, like any other professional in the service industry, and in fact, much like any other worker, will respond to performance measurement when that measurement is deemed to be credible. The attributes of credible data include: (a) a reliable data source that actually reflects the physician's care performance (e.g. measures the care of that physician's patients, not another physician's patients), (b) the measurement of activities that are under the physician's locus of control (e.g. measures the physician's care for the patient's diabetes, not the ophthalmologist's care for the patient's eyes), and (c) reflects current performance not distant past performance. The type of data that are credible to physicians are the ones coming from the medical records of their patients – it's their patient, not someone else's, it reflects the care they have provided, not someone else, and it reflects the most recent care – and from recent patient experience of care surveys, as well as other practice

surveys. These are the data that internists submit to their Board for Maintenance of Certification. Regional Health Information Exchange efforts like the one in Massachusetts are aggregating data from electronic medical records for the purposes of performance assessment. Similar efforts are underway in Cincinnati, Indianapolis and the Hudson Valley in NY. Other efforts such as BTE's have always relied on medical record data to assess performance, and in five years, there has never been a physician that has questioned the credibility of the data or the assessment. As such, assessing performance from medical records is a Best Practice.

b. Best Practices in Measuring Care Efficiency:

- i. Episodes R' Us – Use the most updated versions of the two most common episode analysis packages available, Thomson-Medstat's Episode Groupers, and Ingenix-Symmetry's Episode Treatment Groupers. You can find comparisons of the analyses performed by these programs in the MedPAC and PBGH reports cited earlier. New releases come out regularly and each release does have updates that are important to use. Also use the help of organizations that have performed these analyses before and understand the complexity of the applications and the impact of selecting certain analytical options over others. Clearly both Thomson and Ingenix have depth of experience using their products. Other organizations include Mercer Human Resources, whose analytical team has been analyzing episodes using ETGs for many years.
- ii. The ADS and other pitfalls to avoid – A is for Attribution and it's really important to make the right decisions on how to attribute patients to physicians. We've recommended a one to many approach where patients can be attributed to more than one physician, reflecting the reality of today's patients being often managed by more than one physician. D is for Denominator, and the minimum denominator size is 30. This will likely result in two effects unless the plan is actively participating in an administrative data aggregation effort: (a) only 20% to 30% of individual physicians in the network can be reliably measured, and (b) the base unit of measure is likely to be the group/practice as opposed to the individual physician. S is for Standard error, and any comparative rating of a physician on efficiency will have a standard error and should be calculated with a 95% confidence interval. The CI will create the upper and lower boundaries around the physician's true efficiency performance, and we recommend that the physician always be given the benefit of the doubt in the final ranking. As such, for example, if the physicians are split into quartiles, any physician should be ranked in the quartile in which

the upper bound of the CI is located. Other pitfalls to avoid include creating too much of a hodge-podge of episodes per physician to get to a 30 minimum. There can be significant differences in the average severity of patients in any disease class, and taking all those differences and averaging them out by combining different disease episodes together can mask the actual performance, even with case-mix adjustments. The episode measurement system should also have robust severity adjustments to recognize the fact that more complex patients will require more services. A study in MN on patients with diabetes showed that the average cost of patients with three or more co-morbid conditions was 300% higher than for patients without those co-morbidities. Finally, peer comparisons should include at least 30 physicians within that peer group or else the inter-rating reliability will be poor and the ability to actually rank and distinguish the performance of one physician over another will be very weak.

**C. Pulling It All Together:**

After assessing a score for the effectiveness of the care delivered, and a separate score/ranking for the efficiency of the care delivered, there are several principles that we recommend applying and are summarized in the grid below:



## Incentive grid for “E2” physicians

		Efficiency of Care Ranking			
		Lowest Quartile	Third Quartile	Second Quartile	Highest Quartile
Effectiveness of Care Score	95				
	80	min.			
	65	min.	min.		
	50	min.	min.		

- a. Create a minimum threshold of quality performance under which the physician will not receive any incentives, irrespective of their efficiency ranking, and have clear targets with increasing incentives for achieving high quality

- b. Highly effective care always trumps efficient care because of the uncertainty linked to efficiency measurement and the importance of sending a strong signal to physicians and consumers that high quality care is always rewarded and recognized. As such, effective care above the minimum score should also get a minimum incentive.
- c. Incentives increase as effectiveness and efficiency improve.

A current Best Practice that combines all the elements listed above is CareFirst's Quality Rewards program<sup>10</sup>, and is also the first incentive and reward program in the country that received BTE's Endorsement.

---

<sup>10</sup> <http://www.carefirst.com/providers/html/CareFirstQualityRewards.html>