CME credits available for this article — see page 96.

REVIEW ARTICLE

Pay for Performance for Salaried Health Care Providers: Methodology, Challenges, and Pitfalls

John R Britton, MD

Perm J 2014 Winter;18(1):78-85

http://dx.doi.org/10.7812/TPP/13-087

Abstract

Pay for performance has been recommended by the Institute of Medicine as an incentive to improve the quality of health care. Traditional quality-improvement methods may be adapted to evaluate performance of salaried providers, but it is important to separate provider contributions from other influencing factors within the health care system. Accurate recording, extraction, and analysis of data together with careful selection and measurement of indicators of performance are crucial for meaningful assessment. If appropriate methodology is not used, much time, effort, and money may be expended gathering data that may be potentially misleading or even useless, with the possibility that good performance may go unrecognized and mediocre performance rewarded.

Introduction

Pay for performance has been recommended by the Institute of Medicine as an incentive for physicians to improve the quality of health care.[1] It is based on the notion that financial remuneration of health care providers should be proportionate to the quality of care they provide. Implicit in this concept is the belief that improvement in provider performance will result in improvement in the overall quality of medical care, an intuitively attractive assumption that remains unproven.[2-4] If this concept is correct, it is reasonable to assume that the assessment of provider performance should follow generally accepted principles of quality improvement (QI) currently used in health care settings. Although such principles usually apply to systems or microsystems within a larger system, they may be adapted to evaluate performance of individual providers or groups of providers within a discipline.

Yet, in contrast to institutional QI initiatives, pay-for-performance initiatives have two unique characteristics. First, they must measure the contribution of the providers being assessed independently of the contributions of other components of the health care system, including those of other types of providers. Second, a monetary value must be assigned to the degree of improvement measured, necessitating value judgments that may be somewhat arbitrary, especially when comparing different groups of providers in different settings or specialties.

One of the most important steps in assessing provider performance is the choice of one or more measures of that performance, often referred to as metrics or indicators. If indicators are not carefully chosen, all subsequent improvement efforts may be limited or even useless. Moreover, it must be possible to extract and analyze data relevant to the indicators in an efficient, timely manner. Therefore, much of the discussion that follows in this article will focus on indicator selection, documentation, and measurement.

An extensive literature on pay for performance and its advantages and disadvantages exists, including a number of systematic reviews and case models.[2-5] That information will not be reviewed here. Rather, the methodology, challenges, and pitfalls relevant to the implementation of pay for performance for salaried health care providers within a health care organization will be presented.

Initial Selection of an Indicator

Indicators may relate to any aspect of medical care for which need for improvement is known or perceived based on past performance, variability of care, or adverse outcomes. Although indicators may be chosen for individual providers, they are more often selected by departments within a system for use by all providers within that department. Provider performance may be assessed individually or collectively within a discipline; the latter may be preferable for departments with low patient numbers or substantial shared patient management. The initial choice of an indicator is tentative and its final selection depends on defined need for improvement and feasibility concerns, discussed below.

Candidate measures may be proposed by any individual or group within a department, with subsequent review and selection through the use of focus groups or the department as a whole. Ideally, preexisting consensus-based measures would be chosen when possible, including those suggested by national groups such as the American Medical Association Physician Consortium for Performance Improvement[6] or the Centers for Medicare & Medicaid Services.[7] In addition, measures should be selected for which baseline performance is low, permitting larger effect sizes[5] (see Sample Size Considerations section). Performance indicators should not only have face validity, but improvement associated with them should also be consistent with improvement associated with other indicators that evaluate similar aspects of health care (construct validity). They should also be amenable to consistent measurement (reliability). At the same time that indicators are considered, interventions to improve performance as measured by the indicator should be formulated, together with a tentative plan for their implementation. Literature review should establish an evidence base for the indicator and, if possible, the interventions.

At times, composite indicators may be chosen that combine multiple indicators of quality into an aggregate score. This

John R Britton, MD, is a Neonatologist for the Colorado Permanente Medical Group in Denver, CO. E-mail: john.r.britton@kp.org.

approach is especially attractive if sample size is inadequate for an individual metric but acceptable for the aggregate. Of necessity, this requires assignment of defined weights to each component metric, a process that may be arbitrary. Such an approach may be useful for the tentative analysis of data over short periods, but it obscures the individual performance contributions of the component metrics, allowing superior performance on one component to compensate for poor performance on another.[8]

Structure, Process, and Outcome

Traditional quality measurements have been categorized into three domains[9]: structure (characteristics of the health care environment), process (care administered to the patient), and outcomes (health status of the patient). These domains continue to be emphasized by the Health Resources and Services Administration of the US Department of Health and Human Services.[10]

Structure measures describe features of a health care organization relevant to its capacity to provide health care, usually as manifestations of the setting in which health care is administered and the policies that direct that care. Although such measures may be affected by provider performance, structural measures are poor indicators for assessing individual providers because they also reflect the net result of system factors beyond provider control (confounders).[11] Structural confounders may sometimes be controlled for in analyses, but they are often multiple, vaguely defined, and difficult to quantitatively separate from one another.

Outcomes, such as medical conditions, would appear the most desirable measures of provider performance because they assess patient health status after care and are important in their own right. Yet, although outcomes may sometimes be uniquely attributable to provider performance they, like structure, usually reflect the impact of other confounding factors. In addition, random variation may contribute to outcomes, rendering sample size considerations crucial to interpretation of outcome data. Finally, many outcomes may have low incidence in the patient population studied, making statistical significance difficult to achieve. Several preconditions have been recommended for the selection of outcome indicators, including adjustment for confounders.[2,5,12]

In contrast to structure and outcomes, processes under direct provider control may be more sensitive indicators of individual provider performance.[13,14] Processes are usually not as vulnerable to the effects of confounders and may reflect true changes in provider behavior. Moreover, process indicators are easier to extract from the medical record and interpret, such that data extraction may be less time consuming and statistical analysis more simplified, often requiring only bivariate methods. In contrast to the minimal improvements usually observed for outcome measures, process metrics generally yield better improvement rates.[5,15]

Improvement in an indicator requires selection of an intervention to improve performance. For process indicators, the intervention should lead to improved performance of the process. For outcome indicators, the intervention should facilitate improvement in a process (one or more specific provider behaviors) that will lead to improvement in the outcome. Clearly, analysis of a process indicator will be simpler and more direct than analysis of an outcome indicator. Characteristics of process metrics and an approach to formulating them have been outlined by Rubin et al.[16,17]

For a process to be of value as an indicator it must be linked to an improved outcome,[14,18] and this linkage should be evidence based. This is a very important prerequisite for choice of any process metric, and in its absence, the choice of metric is unjustified.

Confounders

Confounders are factors other than provider performance that may influence the indicator being measured. Unless documented and quantified, these factors may confuse measurement of the indicator with a magnitude that distorts or obscures the contribution of the providers being assessed. Examples of confounders include aspects of case mix, such as demographic and socioeconomic factors, comorbid medical conditions that may affect the indicator, and contributions of other health care providers from other disciplines. Case mix is especially relevant to QI efforts if 1) measures of improvement are related to patient characteristics and 2) the same characteristics differ in their distributions among populations being compared.[19] Case mix characteristics are usually intrinsic to the patient and would not change if the patient were assigned to another population.

To dissect the provider contributions to an indicator, meticulous documentation and measurement of all other contributing factors, often referred to as risk adjustment, is required. This may be difficult, as much of this information is not routinely documented in the medical record and may be time consuming and expensive to obtain. Although direct and indirect standardization methods may correct for confounder effects on indicators, regression methods provide the best means of adjustment.[20] Failure to appropriately adjust for confounders may lead to erroneous results and mistaken conclusions whenever populations are compared, whether the comparison is among clinical units, to national benchmarks, or between groups of patients within the same unit before and after an improvement initiative. If ignored, confounders may obscure good provider performance and compensate for poor provider performance.

Examples of Structural, Process, and Outcome Indicators for Assessment of Provider Performance

1. A Structural Measure with Multiple Confounders

Time to third next available appointment, the average length of time in days between the day a patient makes a request for an appointment and the third next available appointment, is a structural measure of access to care that has been recommended by the Institute for Healthcare Improvement.[21] For example, a dermatology group chooses this indicator as a measure of its performance for the upcoming year after an intervention to increase the number of patients scheduled daily per physician. However, they practice in a rapidly growing health maintenance organization in which membership continues to dramatically increase. In addition, a midlevel provider that conducts group visits for acne patients at the clinic takes family leave for a five-month period, and patients who would have attended this clinic are given appointments with dermatologists. At the end of a year, no reduction in the time to third next available appointment is observed, despite increased patient visits per dermatologist.

As a measure of appointment waiting times, time to third next available appointment is affected by many system factors, only some of which are identifiable. These include appointment types, scheduling patterns, availability of group visits or online consultations, cancellations and no-shows, office hours, support staff, patient characteristics (case mix), number of providers, and provider efficiency. Many of these factors are difficult to quantify, and their contributions to the indicator may change over time. Consequently, physician contributions to the indicator may be obscured, as in the above example. Furthermore, it is possible that reducing the time to third next available appointment by increasing the number of patients seen by a provider in a given time interval may compromise quality of care rather than enhance efficiency. For these reasons, time to third next available appointment, although a valuable measure of the quality of a health care system, is a poor choice for measuring physician performance.

Summary of Pay-for-Performance Methodology

Define major data for uniform documentation within each specialty

- Historical aspects
- Symptoms
- Signs
- Physical findings
- Laboratory data
- Radiography
- Diagnoses
- Therapies
- Outcomes
- Factors influencing outcomes (confounders/risk adjustment)

Data recording and analysis

- Modify electronic medical record to facilitate ongoing discipline-specific improvement efforts
  - Provide fields for structured data
  - Adopt standardized phraseology for unstructured charting (semistructured)
  - Allow customization/addition of new fields for structured data as needs arise
  - Develop temporal tags for structured entries
- Facilitate
  - Simple, rapid query of structured and semistructured data by providers
  - Export of query results into database
  - Analysis of database data with statistical software

Indicator selection

- Process measures with evidence-based link to favorable outcomes
- Proposed by individuals and/or focus groups
- Applicable for provider groups/departments
- Review literature to confirm evidence base
- Measurable
- Valid
- Reliable
- Suboptimal past performance or new practice
- Confounders defined and measurable (risk adjustment)
- Sample size adequacy (number of patients, time periods for analyses)
- Define anticipated magnitude of meaningful improvement

Implementation of practice change during interim period

Interventions to enhance performance

Continuous performance evaluation and feedback

- Run/control charts
- Bivariate statistics for before-and-after comparisons
- Multivariate techniques to control for confounders

Define relationship between magnitude of improvement and remuneration

Other structural metrics frequently chosen to assess QI include reduction of door-to-balloon time (the interval between arrival in the emergency room and cardiac catheterization for a patient with myocardial infarction), enhanced operating room block use, reduction of hospital readmission rates, and improved Emergency Department use. These are all excellent improvement initiatives for a hospital or health care system, and in some cases it may be feasible to measure and control for nonprovider contributions to the outcome, such that provider performance can be independently assessed. However, in most cases this is very difficult if not impossible to accomplish because of multiple, poorly defined confounders, rendering these structural indicators unfavorable measures of provider performance.

2. An Outcome Indicator with Limited Confounders

Vitreous loss during cataract surgery is a complication that occurs in approximately 8% of surgeries.[22] As an example, a group of ophthalmologists observes that their rate of vitreous loss is 16% and embarks on an initiative to improve performance during the upcoming year. After literature review, they determine that the incidence of vitreous loss is dependent on surgeon experience, patient volume, and case complexity.[22] Patient volume in their practice has been fairly constant over several years and they project similar patient numbers for the future. In addition, they have been using a scoring system that provides for preoperative risk stratification of individual cases.[22] They decide to continue to use this score to control for patient complexity in future measurements of surgeon performance. They undertake an educational initiative that includes group review of surgical technique and visits to another practice with a low rate of vitreous loss, after which they begin a year of measurement, with ongoing group review and discussion of each occurrence of this complication. At the end of the year, after risk adjustment, they observe a statistically significant decrease in vitreous loss, with an incidence of 11%.

This is a reliable and valid outcome indicator to which limited, definable, and measurable factors contribute. The risk scoring system provides for ongoing case mix adjustment, allowing the contribution of surgeon performance to be evaluated with reasonable accuracy.

3. A Process Measure with Proven Links to Favorable Outcomes

Prenatal steroid treatment for threatened preterm birth has been shown to substantially reduce a variety of morbidities such as respiratory distress syndrome among very low-birth-weight infants. The rate of prenatal steroid treatment is used as a benchmark for the quality of perinatal care by the Vermont Oxford Network, an international collaborative.[23] Reviewing the literature at a department meeting and recognizing that the rate of prenatal steroid treatment among their patients has been low for the past year, a group of perinatologists and obstetricians decides to embark on an initiative to improve prenatal steroid administration to eligible patients in the coming year, and a portion of their salary will be based on the degree of improvement. At the end of the year, they observe a significant increase in the rate of steroid administration. However, the incidence of morbidity among very low-birth-weight infants does not change.

Prenatal steroid treatment is a process that the literature has clearly linked to favorable outcomes among very low-birth-weight infants, and because the process is exclusively directed by the perinatologist or obstetrician, it is a reflection of physician performance. Importantly, neither case mix nor other providers substantially influence this process. Such factors may have influenced the neonatal outcomes in the above example, given that they remained unchanged. Although rapid progression of labor with delivery before steroid administration (as might occur with patients who do not seek timely care) could result in reduced apparent performance rates for steroid administration, these cases could be controlled for in analysis. In addition, available benchmark data from the Vermont Oxford Network provide an attractive added benefit.

Provider prescribing patterns are among the best process measures of performance because they are usually linked to improved outcomes, reflect the direct actions of the provider, and are relatively easy to extract from electronic medical records (EMRs). Another useful process measure of provider performance is radiologist report turnaround time, which has an impact on the efficiency of care and has been shown to be affected by pay-for-performance efforts.[24] Meaningful use measures, such as maintenance of patient problem, medication, and allergy lists, might be other useful process metrics for similar reasons.[7]

Improvement of Past Performance: Are Relevant Data Available?

For established health care practices, choice of an indicator is often based on known or perceived suboptimal past performance for that practice. Although subjective impressions or adverse events may prompt proposed indicator candidacy, suboptimal performance should be confirmed by retrospective analysis of performance data. Unfortunately, such data are frequently unavailable and their extraction may require labor-intensive, time-consuming review of medical records. As a result, demonstration of suboptimal past performance often poses a major hurdle that must be overcome if further improvement efforts are to proceed.

Sometimes an indicator may involve the adoption of a new practice that is evidence based and holds promise to provide for potentially better care (potentially better practice).[25] Because the practice is new, there are usually no data on past performance that can be used to gauge need for improvement. Often it may be possible to obtain data from other groups, such as community samples, or published performance rates as a benchmark. Calculation of the required sample size (see Sample Size Considerations section) becomes problematic, and estimates of target parameters for improvement may be arbitrary. As a result, there is great potential for either underestimating or overestimating performance targets. An advantage, however, is that with no past performance, even modest improvement may be easy to achieve.

Structured Data in Medical Records

To measure provider performance requires data, which usually must be extracted from medical records. Although such data may be extracted manually from traditional paper records, the EMR has gained increased use in hospitals and physician offices since the Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009.[26,27] Because it is amenable to computerized query, the EMR has great potential to serve as an ongoing source of data for past and current provider performance measures.[26,28] However, the information that can be extracted from an EMR is only as good as that which is entered.[29] Consistency of data entry among providers is crucial if accurate data are to be obtained and compared.

Data are most amenable to extraction if they are entered into dedicated or "structured" fields in an EMR.[30] Although data such as vital signs, demographic characteristics, medications, and diagnosis codes may be contained within such fields, much information relevant to provider performance is often embedded in unstructured, free-text notes. Information in such form may be influenced by variability in provider charting style and subjectivity of assessment.[31] Because it may be positioned randomly within a note, it must be extracted manually or by sophisticated computer programs that are not in general use. At best, this is a laborious process fraught with potential error, due in part to the multiple variations in documentation style among providers. Before embarking on an initiative to improve performance, it is important to determine which variables must be extracted from the EMR to analyze performance, whether they exist in structured or unstructured form, and the degree of effort that will be required for their extraction.

… information that can be extracted from an EMR is only as good as that which is entered.[29] Consistency of data entry among providers is crucial if accurate data are to be obtained and compared.

For data to reflect past performance they must already exist within the EMR. Here lies the problem: To identify areas in need of improvement we need to know past performance, but to know past performance, we need to have appropriate, prospectively recorded data to measure that performance in a consistent, easily extractable form in the EMR. This may be difficult to achieve given the current predominance of unstructured provider charting in the EMR. Quality measures based on unstructured fields are likely to be inaccurate.[32] Changing provider documentation styles to achieve consistency and enhanced use of structured charting may be needed.

Within any medical specialty there is a body of discrete data for which fields could be generated within an EMR for provider use. These could take the form of separate pull-down entries or standard phrases that could be consistently used by providers within that specialty in a given patient care situation. For inpatient care, temporal linkage of structured data would be helpful for future analysis (eg, time since admission, surgery, delivery, rather than absolute date and time). If this could be achieved, retrospective data extraction for performance analysis could be easier.

Sample Size Considerations

A frequently overlooked aspect of performance analysis is adequacy of the size of the analyzed sample of a population. If sample size is inadequate, there is a substantial risk of concluding that there was no change in performance when, in reality, a change did occur.[19] This phenomenon is referred to as a type II error, and its probability is symbolized by the Greek letter β. Statistical power (1 - β) is the probability of correctly rejecting a false null hypothesis and detecting an effect that truly exists. Before any measurement of performance is made, the minimum adequate sample size should be determined. It is always desirable to maximize power, and typically a value of 0.8 or greater is chosen.

Statistical power is primarily determined by four factors: α level, directionality of measurement, effect size, and sample size.[33]

1) The α level, traditionally symbolized by the letter p, is the probability of rejecting the null hypothesis when it is true, thereby concluding that a difference is present when it is not. Most frequently, a p value of 0.05 or less is chosen as an acceptable α level.

2) Directionality of measurement is implicit in the process of QI initiatives, given that we usually hypothesize that our interventions will improve, rather than worsen, performance metrics. As a result, we may confidently use one-tailed, rather than two-tailed, statistical tests in our analyses.

3) Effect size assesses the anticipated magnitude of change with respect to a given metric. The effect size is usually not under our direct control when we embark on an improvement initiative, because we cannot predict how much of an improvement will occur. We might estimate effect size on the basis of past changes in our population or on the basis of improvement results reported by others (benchmarking). And although effect size may be arbitrarily estimated, its magnitude will have practical implications only within a particular clinical context. Lower baseline performance may permit greater room for improvement and a greater effect size.[5]

4) Of the factors that influence statistical power, sample size has the greatest impact. With increasing sample size, the standard error of the distribution sampled is progressively reduced. Adequacy of sample size should always be determined before beginning any improvement initiative.

Sample size calculations will depend on the statistical test chosen for comparative analysis. Choice of an inappropriate statistical test for comparison may influence the power. For example, applying parametric rather than nonparametric tests to nonnormally distributed data reduces power and increases the probability of a type II error, a common mistake in the analysis of Emergency Department length-of-stay data.[34]

After choosing the appropriate statistical test to compare the outcome of interest before and after the initiative, it is reasonable to choose the one-sided version of that test and set the α and β levels to 0.05 and 0.20, respectively (80% power). Once these parameters are set, the magnitude of the observable effect will be determined by the sample size available for analysis. Generally, larger numbers of subjects are required to detect smaller changes in performance.

Several Web sites are available that permit determination of the sample size required to achieve a desired effect size.[35-37] Sample size should be determined a priori, but power may also be assessed periodically during the accumulation of data, permitting termination of study when an adequate sample size is achieved.

The Importance of Sample Size Determination: An Example

As discussed above, vitreous loss during cataract surgery is an attractive outcome measure that may reflect surgeon performance after case mix adjustment. Consider a salaried ophthalmologist, who finds that the incidence of this complication among her 250 cataract patients was 16% during the past year, much higher than the optimal value of 8% reported in the literature.[22] Further analysis reveals that case mix of her patients has been stable with time. Accordingly, she pursues additional training and experience in cataract surgery and chooses this metric to assess her performance for the upcoming year. Discussions with her supervisors lead to the conclusion that she will receive a monetary bonus if she can reduce the rate to 10% during the coming year. After performing 250 cataract surgeries that year, her rate of vitreous loss is 13%, a value deemed not significantly different from the previous year by a χ^2 test (p = 0.13). As a result, she does not receive the bonus.

If a sample size assessment had been made at the onset of this initiative, it would have been determined that comparing the outcome in 250 patients during the "test" year (after the intervention) to that for an identically sized group during the previous year would yield a power of 0.5. This means that there was a 50% chance of not detecting an improvement even if one had truly occurred. At the ophthalmologist's current rate of cataract surgeries, it would take 2 years to treat enough patients (500 to compare to a similar group for the previous 2 years) to attain a power of 0.80. Thus, it is entirely possible that the surgeon was performing at a higher level but that her improvement went undetected, with consequent loss of remuneration.

Databases and Statistical Programs

Ideally, data fields for quality initiatives and pay-for-performance measures would be integrated into evolving EMR systems as the latter are developed instead of retrospectively, a more difficult task.[32] The EMR populations should be amenable to rapid and efficient query by members of the QI team to obtain information needed for performance analysis. The results of such queries should be exportable into database programs. However, this may not be possible, especially if the EMR lacks the required structured data fields or is not amenable to easy query. In cases such as this, a separate database into which data may be entered manually may be required.

After population data are entered into a database, preliminary analyses may often be performed within that database. Some database programs permit performance of elementary tasks, such as sorting, variable definition, and bivariate statistical analyses. However, more sophisticated statistical procedures, such as multivariate analyses, are best performed with a statistical program that can import data from database programs.

Choice and Implementation of the Intervention

The intervention is essentially a prompt to foster performance improvement. It may take many forms and will attempt to facilitate a change in provider practice that either constitutes (process) or affects (outcome) the indicator. Given that the providers will have selected the indicator as a measure of performance, they will presumably be keenly aware of the need to improve and of the financial incentives for doing so. It is assumed that this awareness will be adequate to implement the performance change. However, educational interventions in the form of formal presentations, posters, and message prompts may help to maintain this awareness. If analysis of past performance, ongoing monitoring, or both reveal suboptimal performance for certain individuals within a department, feedback could be provided to those providers. Periodic review of performance could also be presented at regular department meetings, with case review as needed. A method for implementation of practice change was previously reported.[38]

Ongoing Monitoring of Performance

Performance improvement is best assessed by comparison of sequential time periods before and after the implementation of an initiative using bivariate and multivariate statistics as appropriate. In addition, control charts may be useful in determining ongoing trends during an improvement initiative, with continuous feedback to providers and ongoing review of cases. Because of the time required to implement practice changes, it may be advisable to allow for an interim period between the periods to be compared, during which time changes may be made. Use of moving averages or rolling period analysis may add precision when samples are small during individual time periods; however, this approach is insensitive to changes from one period to the next.[19] Moreover, there is a middle-period bias, such that higher performance in the middle period will yield a higher rolling period performance rate than will performance at either end period. Thus, more recent improvements may go unrecognized.

Incentives and Meaningful Magnitude of Improvement

For general QI initiatives, any improvement that is statistically significant and clinically meaningful may be considered important. However, translating the magnitude of improvement into monetary compensation becomes a challenge. Inevitably, somewhat arbitrary value judgments must be made in such assignments, especially when comparing different metrics among departments or specialties. Moreover, all metrics may not be amenable to similar degrees of improvement, even with the best efforts on the part of individual providers. Providers with poor past performance have greater potential for improvement, in contrast to high-performance providers with less room for improvement.

Most frequently, improvement incentives consist of financial rewards that may be based on either a defined threshold for improvement or a continuous scale. In general, they produce a greater positive effect for low performers than for high performers.[5] However, the relationship between incentive size and improvement effect has not been established. The frequency of incentive payment has been reported not to affect performance.[39]

Statistically Significant Change Versus Meaningful Change: The Example of Patient Satisfaction

It is important to be aware that all statistically significant improvements may not be clinically meaningful because of their magnitude. A good example of this phenomenon is the measurement of patient satisfaction. Although few studies have demonstrated that patient satisfaction is associated with quality of provider care,[40] satisfaction surveys have been recommended as a quality indicator by the Institute of Medicine[1] and have been adopted by many hospitals and practice groups. Consider the following example: A large multidisciplinary group practice uses satisfaction surveys to assess the quality of physician care during recent office visits and hospitalizations. These surveys yield scores that range from 0% to 100%. The practice has determined that scores on the surveys will partially determine physician bonus payments within departments. During sequential 12-month periods, scores for medical subspecialty care increased from 94% to 96%. However, scores for surgical subspecialties declined from 94% to 93%. Both of these changes were statistically significant, with $p < 0.01$.

Studies of patient satisfaction tend to show high levels of undifferentiated satisfaction, with most respondents rating the quality of provider care very high,[40] even when assessed across multiple categories. One consequence is that large patient samples are required to detect significant changes in performance ratings, because scores may already be near the top of the measurement scale, with little room for improvement. Even if significant changes are observed, the magnitude of change is likely to be small and of questionable practical importance, as shown in the above example. It has been suggested that satisfaction surveys should attempt to focus on dissatisfaction with care rather than satisfaction.[41] If results of satisfaction surveys with high levels of satisfaction are chosen as a basis for provider remuneration, it might be preferable to reward sustained high performance above a chosen threshold, rather than small but statistically significant changes of questionable practical importance.

Reward for Development of Indicators and Analysis of Performance

Metric selection and assessment of feasibility may be time consuming and complicated. A considerable amount of effort may be required to extract data on past performance, review relevant literature, modify charting methods, and set up databases to prospectively monitor improvement. In some cases it may be reasonable to provide compensation for these efforts, yet the magnitude of compensation for particular levels of progress will remain arbitrary. In addition, some initially attractive metrics may later prove impractical for unforeseen reasons, but only after the expenditure of considerable effort to explore their feasibility. The possibility of such results must be anticipated and provisions made for appropriate remuneration for efforts expended.

Even when appropriate indicators are chosen and optimal computer facilities are available for documentation and analysis, a substantial amount of time is required to complete an assessment of performance. One or more individuals comfortable with the methodology described above should be designated to coordinate these efforts, and the cost of their financial remuneration should be anticipated at the onset.

Summary and Conclusions

Most aspects of pay for performance are not unique but are shared by QI initiatives in general (see Sidebar: Summary of Pay-for-Performance Methodology). Achieving uniform consensus-based, specialty-specific documentation in the EMR is a worthy goal in its own right. An especially important aspect that has not received adequate attention in the literature is the need for EMRs that are more conducive to data collection for quality purposes, with structured charting and capacity for rapid query by improvement teams. Current EMRs are deficient in these attributes, and investment of time and resources to rectify these deficiencies would greatly enhance the capability and efficiency of all future QI efforts.

In pay for performance, process indicators with known links to favorable outcomes are preferred to outcome or structural indicators, and a priori determination of sample size adequacy is crucial if erroneous conclusions are to be avoided. Continuous evaluation of performance with ongoing feedback to providers during the initiative is also critical. Yet even if improvement can be rigorously accomplished, achievement of equitable remuneration among provider groups will remain a challenge because of the arbitrariness inherent in assigning monetary value to the degree of improvement.

The methodologic goals outlined above are neither esoteric nor complex. They are basic, achievable, and important for accurately conducting provider-specific QI initiatives. The technologies required to implement them are simple and inexpensive in any setting with EMRs and a personal computer. If this methodology is ignored, much time, effort, and money may be expended in gathering data that may be potentially misleading or even useless, such that good performance may go unrecognized and mediocre performance rewarded. ❖

Disclosure Statement

The author(s) have no conflicts of interest to disclose.

Acknowledgment

Leslie Parker, ELS, provided editorial assistance.

References

1. Institute of Medicine. C1. Committee on Quality Health Care in America, Institute of Medicine. Crossing the quality chasm: a new health system for the 21st century. Washington, DC: National Academies Press; 2001 Mar 1.
2. Petersen LA, Woodard LD, Urech T, Daw C, Sookanan S. Does pay-for-performance improve the quality of health care? Ann Intern Med 2006 Aug 15;145(4):265-72. DOI: http://dx.doi.org/10.7326/0003-4819-145-4-200608150-00006.
3. Scott I. What are the most effective strategies for improving quality and safety of health care? Intern Med J 2009 Jun;39(6):389-400. DOI: http://dx.doi.org/10.1111/j.1445-5994.2008.01798.x.
4. Scott A, Sivey P, Ait Ouakrim D, et al. The effect of financial incentives on the quality of health care provided by primary care physicians. Cochrane Database Syst Rev 2011 Sep 7;(9):CD008451. DOI: http://dx.doi.org/10.1002/14651858.CD008451.
5. Van Herck P, De Smedt D, Annemans L, Remmen R, Rosenthal MB, Sermeus W. Systematic review: effects, design choices, and context of pay-for-performance in health care. BMC Health Serv Res 2010 Aug 23;10:247. DOI: http://dx.doi.org/10.1186/1472-6963-10-247.
6. Physician Consortium for Performance Improvement [Internet]. Chicago, IL: American Medical Association; c2013 [cited 2013 Aug 10]. Available from: www.ama-assn.org/go/pcpi.
7. Centers for Medicare & Medicaid Services [Internet]. Baltimore, MD: Centers for Medicare & Medicaid Services; c2013 [cited 2013 Aug 10]. Available from: www.cms.gov.
8. Profit J, Typpo KV, Hysong SJ, Woodard LD, Kallen MA, Petersen LA. Improving benchmarking by using an explicit framework for the development of composite indicators: an example using pediatric quality of care. Implement Sci 2010 Feb 9;5:13. DOI: http://dx.doi.org/10.1186/1748-5908-5-13.
9. Donabedian A. Evaluating the quality of medical care. Milbank Meml Fund Q 1966 Jul;44(3):Suppl:166-206. DOI: http://dx.doi.org/10.2307/3348969.
10. What sorts of metrics assess quality improvement? [Internet]. Rockville, MD: Health Information Technology and Quality Improvement, US Department of Health and Human Services; c2013 [cited 2013 Jan 26]. Available from: www.hrsa.gov/healthit/toolbox/RuralHealthIT-toolbox/PatientQuality/metrics.html.
11. Birkmeyer JD, Dimick JB, Birkmeyer NJ. Measuring the quality of surgical care: structure, process, or outcomes? J Am Coll Surg 2004 Apr;198(4):626-32. DOI: http://dx.doi.org/10.1016/j.jamcollsurg.2003.11.017.
12. Conrad DA, Perry L. Quality-based financial incentives in health care: can we improve quality by paying for it? Ann Rev Public Health 2009;30:357-71. DOI: http://dx.doi.org/10.1146/annurev.publhealth.031308.100243.
13. Mant J, Hicks N. Detecting differences in quality of care: the sensitivity of measures of process and outcome in treating acute myocardial infarction. BMJ 1995 Sep 23;311(7008):793-6. DOI: http://dx.doi.org/10.1136/bmj.311.7008.793.
14. Mant J. Process versus outcome indicators in the assessment of quality of health care. Int J Qual Health Care 2001 Dec;13(6):475-80. DOI: http://dx.doi.org/10.1093/intqhc/13.6.475.
15. Hart-Hester S, Jones W, Watzlaf VJ, et al. Impact of creating a pay for quality improvement (P4QI) incentive program on healthcare disparity: leveraging HIT in rural hospitals and small physician offices. Perspect Health Inf Manag 2008;5:14.
16. Rubin HR, Pronovost P, Diette GB. From a process of care to a measure: the development and testing of a quality indicator. Int J Qual Health Care 2001 Dec;13(6):489-96. DOI: http://dx.doi.org/10.1093/intqhc/13.6.489.
17. Rubin HR, Pronovost P, Diette GB. The advantages and disadvantages of process-based measures of health care quality. Int J Qual Health Care 2001 Dec;13(6):469-74. DOI: http://dx.doi.org/10.1093/intqhc/13.6.469.
18. Hammermeister KE, Shroyer AL, Sethi GK, Grover FL. Why it is important to demonstrate linkages between outcomes of care and processes and structures of care. Med Care 1995 Oct;33(10 Suppl):OS5-16. DOI: http://dx.doi.org/10.1097/00005650-199510001-00002.
19. Zaslavsky AM. Statistical issues in reporting quality data: small samples and casemix variation. Int J Qual Health Care 2001 Dec;13(6):481-8. DOI: http://dx.doi.org/10.1093/intqhc/13.6.481.
20. Burgess JF Jr, Christiansen CL, Michalak SE, Morris CN. Medical profiling: improving standards and risk adjustments using hierarchical models. J Health Econ 2000 May;19(3):291-309. DOI: http://dx.doi.org/10.1016/S0167-6296(99)00034-X.
21. Third next available appointment [Internet]. Cambridge, MA: Institute for Healthcare Improvement; c2013 [cited 2013 Apr 8]. Available from: www.ihi.org/knowledge/Pages/Measures/ThirdNextAvailableAppointment.aspx.
22. Jacobs PM. Vitreous loss during cataract surgery: prevention and optimal management. Eye (Lond) 2008 Oct;22(10):1286-9. DOI: http://dx.doi.org/10.1038/eye.2008.22.

Erratum in: Eye 2008 Oct;22(10):1370. DOI: http://dx.doi.org/10.1038/eye.2008.138.

23. Horbar JD, Soll RF, Edwards WH. The Vermont Oxford Network: a community of practice. Clin Perinatol 2010 Mar;37(1):29-47. DOI: http://dx.doi.org/10.1016/j.clp.2010.01.003.
24. Boland GW, Halpern EF, Gazelle GS. Radiologist report turnaround time: impact of pay-for-performance measures. AJR Am J Roentgenol 2010 Sep;195(3):707-11. DOI: http://dx.doi.org/10.2214/AJR.09.4164.
25. Plsek PE. Quality improvement methods in clinical medicine. Pediatrics 1999 Jan;103(1 Suppl E):203-14.
26. Chan KS, Fowles JB, Weiner JP. Review: electronic health records and the reliability and validity of quality measures: a review of the literature. Med Care Res Rev 2010 Oct;67(5):503-27. DOI: http://dx.doi.org/10.1177/1077558709359007.
27. Blumenthal D, Tavenner M. The "meaningful use" regulation for electronic health records. New Engl J Med 2010 Aug 5;363(6):501-4. DOI: http://dx.doi.org/10.1056/NEJMp1006114.
28. Restuccia JD, Cohen AB, Horwitt JN, Shwartz M. Hospital implementation of health information technology and quality of care: are they related? BMC Med Inform Decis Mak 2012 Sep 27;12:109. DOI: http://dx.doi.org/10.1186/1472-6947-12-109.
29. Menachemi N, Collum TH. Benefits and drawbacks of electronic health record systems. Risk Manag Healthc Policy 2011;4:47-55. DOI: http://dx.doi.org/10.2147/RMHP.S12985.
30. Henry SB, Morris JA, Holzemer WL. Using structured text and templates to capture health status outcomes in the electronic health record. Jt Comm J Qual Improv 1997 Dec;23(12):667-77.
31. Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? J Biomed Inform 2009 Oct;42(5):760-72. DOI: http://dx.doi.org/10.1016/j.jbi.2009.08.007.
32. Weiner JP, Fowles JB, Chan KS. New paradigms for measuring clinical performance using electronic health records. Int J Qual Health Care 2012 Jun;24(3):200-5. DOI: http://dx.doi.org/10.1093/intqhc/mzs011.
33. Statistical power [Internet]. Independence, KY: Wadsworth Cengage Learning; c2005 [cited 2013 Jun 1]. Available from: www.wadsworth.com/psychology_d/templates/student_resources/workshops/stat_workshp/statpower/statpower_04.html.
34. Qualls M, Pallin DJ, Schuur JD. Parametric versus nonparametric statistical tests: the length of stay example. Acad Emerg Med 2010 Oct;17(10):1113-21. DOI: http://dx.doi.org/10.1111/j.1553-2712.2010.00874.x.
35. Brant R. Inference for proportions: comparing two independent samples [Internet]. British Columbia, Canada: Rollin Brant; [cited 2013 Jun 1]. Available from: www.stat.ubc.ca/~rollin/stats/ssize/b2.html.
36. Researcher's toolkit [Internet]. Fort Worth, TX: DSS Research; c2013 [cited 2013 Jun 1]. Available from: www.dssresearch.com/KnowledgeCenter/toolkitcalculators/statisticalpowercalculators.aspx.
37. Soper D. A-priori sample size calculator for multiple regression [Internet]. Fullerton, CA: Daniel Soper; c2013 [cited 2013 Jun 1]. Available from: www.danielsoper.com/statcalc3/calc.aspx?id=1.
38. Pantoja AF, Britton JR. An evidence-based, multidisciplinary process for implementation of potentially better practices using a computerized medical record. Int J Qual Health Care 2011 Jun;23(3):309-16. DOI: http://dx.doi.org/10.1093/intqhc/mzr012.
39. Chung S, Palaniappan L, Wong E, Rubin H, Luft H. Does the frequency of pay-for-performance payment matter?—Experience from a randomized trial. Health Serv Res 2010 Apr;45(2):553-64. DOI: http://dx.doi.org/10.1111/j.1475-6773.2009.01072.x.
40. Britton JR. The assessment of satisfaction with care in the perinatal period. J Psychosom Obstet Gynaecol 2012 Jun;33(2):37-44. DOI: http://dx.doi.org/10.3109/0167482X.2012.658464.
41. Coyle J. Understanding dissatisfied users: developing a framework for comprehending criticisms of health care work. J Adv Nurs 1999 Sep;30(3):723-31. DOI: http://dx.doi.org/10.1046/j.1365-2648.1999.01137.x.

A Unique Service

No greater opportunity, responsibility, or obligation can fall to the lot of a human being than to become a physician. In the care of the suffering he needs technical skill, scientific knowledge, and human understanding. He who uses these with courage, with humility, and with wisdom will provide a unique service for his fellow man and will build an enduring edifice of character within himself. The physician should ask of his destiny no more than this; he should be content with no less.

—Tinsley R Harrison, 1900-1978, American physician and editor of the first five editions of *Harrison's Principles of Internal Medicine*