Published by Oxford University Press in association with The London School of Hygiene and Tropical Medicine *Health Policy and Planning* 2009;**24**:167–174
 Advance Access publication 18 February 2009 doi:10.1093/heapol/czp003

A novel method for measuring health care system performance: experience from QIDS in the Philippines

Orville Solon,[1] Kimberly Woo,[2] Stella A Quimbo,[1] Riti Shimkhada,[2] Jhiedon Florentino[1] and John W Peabody[2]*

Accepted 3 December 2008

Objectives Measuring and monitoring health system performance is important albeit controversial. Technical, logistic and financial challenges are formidable. We introduced a system of measurement, which we call Q*, to measure the quality of hospital clinical performance across a range of facilities. This paper describes how Q* was developed, implemented in hospitals in the Philippines and how it compares with typical measures.

Methods Q* consists of measures of clinical performance, patient satisfaction and volume of physician services. We evaluate Q* using experimental data from the Quality Improvement Demonstration Study (QIDS), a randomized policy experiment. We determined its responsiveness over time and to changes in structural measures such as staffing and supplies. We also examined the operational costs of implementing Q*.

Results Q* was sustainable, minimally disruptive and readily grafted into existing routines in 30 hospitals in 10 provinces semi-annually for a period of 2½ years. We found Q* to be more responsive to immediate impacts of policy change than standard structural measures. The operational costs totalled US$2133 or US$305 per assessment per site.

Conclusion Q* appears to be an achievable assessment tool that is a comprehensive and responsive measure of system level quality at a limited cost in resource-poor settings.

Keywords Quality of care, health systems, health facilities, health policy, developing countries, Philippines, performance measures

[1] University of the Philippines, School of Economics, Diliman, Philippines.
[2] Institute for Global Health, University of California San Francisco, USA.
* Corresponding author. John W Peabody, MD, PhD, DTM&H, Institute for Global Health, University of California, San Francisco, 50 Beale Street, Suite 1200, San Francisco, CA 94105, USA. Tel: +1 415–597–8200. Fax: +1 415–597–8299. E-mail: Peabody@psg.ucsf.edu

KEY MESSAGES

- Measuring health care system performance is a crucial element in improving health systems in developing countries.
- We introduce a measure, Q*, that integrates quality of clinical care into an overall performance measure that we developed and implemented in hospitals in the Philippines.
- Q* is more responsive to immediate impacts of policy change than structural measures, which are typically used in performance evaluation of hospitals.
- Q* was easily implemented as it is minimally disruptive and readily grafted into existing routines of hospitals, and was inexpensive with respect to data collection, computation and feedback.
- Q* offers a comprehensive and effective measurement of quality of care that can be introduced into resource-poor settings.

Introduction

In a developing country setting, measuring health sector performance can be difficult, costly and controversial (World Health Organization 2000). One reason is the dearth of data and concerns about the reliability of the data that do exist. A second concern is that data collection is not done serially or consistently, hence there is little information on trends. These significant shortcomings not withstanding, another major challenge centres on whether performance measurement is feasible and can be readily undertaken effectively, in a transparent manner, by agencies and bureaux that are over-worked, locked into old routines, and resistant to changes in standard practices (Murray and Frenk 2006). Questions about data reliability and lack of transparency in design and measurement make the use of performance indicators not only problematic but also highly contested (Almeida *et al*. 2001). Once the debate becomes political, the value of performance measures is diminished by questions of motive, intention and agenda (Braveman *et al*. 2001). Yet measuring health care system performance is crucial in developing countries if health systems are to improve.

Evaluations of health system performance, in rich or poor countries alike, have only recently begun to include a critical measurement of system performance: the quality of health care (Arah *et al*. 2003; Peabody *et al*. 2006). Improving system performance and quality of care is of particular importance because ultimately these systems improve health by providing effective services. When quality is measured at all, it is done in terms of health care inputs and facility-level characteristics, which are referred to as structural measures (Peabody *et al*. 1994). Structural indicators, for example, describe availability of drugs, supplies and technology, available health manpower (Mainz 2003). The underlying assumption is that with proper settings and instrumentalities, good services will follow, but, obviously, this is not always true (Donabedian 2005). While these structural measures are routinely collected, they are limited because of their indirect and limited impact on health outcomes (Donabedian 1988). Moreover, since structural inputs tend to remain fixed over long periods, they are of limited practical use in tracking how policy initiatives affect day-to-day clinical practices that lead directly to changes in health status.

In recent years there has been a growing interest in measuring the process of care—what happens when the provider sees the patient—and in assessing the quality of clinical care services. This interest stems from the drive to apply evidence-based medicine and the search for cost-effective ways to improve health care (Mainz 2003). Not surprisingly, these initiatives focus on evaluating individual provider performance, and while revealing and effective at improving one provider's clinical practice, they do not capture other measures of performance that are of interest at the system level. Pay for performance, which rewards providers for high quality of care, may be particularly beneficial in the developing country context (McNamara 2006; Soeters *et al*. 2007) where quality of care is low (Peabody and Liu 2007) and improvement has become a focus of attention (Bouchet *et al*. 2002; Ovretveit 2004). In these settings, bonus programmes can encourage critical improvements in health care delivery; for example, in Haiti, paying organizations based on health targets led to significant increases in immunization coverage and attended deliveries (McNamara 2005).

In this paper we report on a system for measuring that integrates quality of clinical care, patient satisfaction and volume of physician services into an overall performance measure, which we call Q*. We describe how we developed the metric and how it was implemented over a large region in the Philippines. In the results section, we report our findings from Q* implementation and then compare Q* with existing structural measures. In the discussion, we consider how it might be applied in other resource-poor settings to monitor the quality of clinical care as one aspect of system performance assessment.

Methods

Setting and funding

The Quality Improvement Demonstration Study (QIDS) is a 5-year project begun in 2003 in the Philippines, under the aegis of a unique partnership between the Philippine Health Insurance Corporation (PHIC), the Philippines Department of Health, the University of California San Francisco, and the UPecon Foundation at the University of the Philippines School of Economics. PHIC is the country's social health insurance programme and the largest third party payer for inpatient care. The project is funded by NICHD R-01 #HD042117 and registered with ClinicalTrials.gov. The study was conducted in accordance with the ethical standards of the applicable

national and institutional review boards (IRBs) of the University of the Philippines and the University of California, San Francisco.

QIDS encompasses 30 district hospitals in 11 provinces in the Visayas and Northern Mindanao, the central regions of the Philippines. The catchment areas of the 30 selected hospitals contain approximately 1 million households.

Study design

QIDS uses an experimental design with randomization of two interventions and a control group to evaluate the impact of two major health sector reform initiatives on the health status and cognitive development of children aged 6 to 59 months (Shimkhada *et al*. 2008). Hospital sites in the design were matched and grouped into blocks of three based on population characteristics such as average income and percentage of population with insurance, as well as system characteristics such as number of specialists and proximity to Manila. Each site was randomly assigned to one of the two interventions or the control group. The study collected baseline data in Round 1 and follow-up data in Round 2 to evaluate the differences between groups over time. Interim data were also collected every 6 months from providers and facilities to provide intercurrent information that might change rapidly over time.

The QIDS Bonus Intervention introduced a quality of service (performance) based payment for hospitals and physicians. If the quality of care provided by a group of physicians in one of the 10 Bonus facilities met a specified threshold, then a financial reward, paid by PHIC, was an additional 100 pesos per day of confinement (US$2.15). This implied a 6–18% increase in annual physician income. Ten QIDS sites were randomly assigned to the Control group (C sites). At C sites, PHIC policies and practices are followed as normally would be the case without any special interventions. The policy impacts of the Bonus Intervention (henceforth called Intervention) relative to Control sites were monitored using Q* and other structural measures of quality.

Data for this study were collected between December 2004 and December 2006 by QIDS staff. Three surveys were used to collect the data and are described in detail below.

Facility and mini facility surveys

The facility survey collected comprehensive data on a broad range of facility characteristics that include structural variables, patient case load, demographics, staffing, and other key variables, such as costs and availability of services to area households. Due to the extensiveness of information collected from the facility survey, it was administered once during baseline and again during 24 months post-intervention in Round 2. An abbreviated version of the facility survey (called the mini facility survey) tracked various structural variables and utilization measures quarterly.

After a careful review of the literature and assessment of the local policy requirements, a limited number of key structural measures used to evaluate quality of care were selected for inclusion in our facility survey (Peabody *et al*. 1998). The collected structural variables included the following: (1) number of doctors and nurses; (2) number of prescriptions filled in the past week; (3) number of laboratory tests available; (4) availability of supplies; (5) number of functioning items of equipment; and (6) number of functioning medical instruments. Variables that vary over the short term, such as number of prescriptions filled, availability of laboratory tests and medical supplies, are monitored using mini facility surveys. The structural variables that are not likely to change in the short term, such as medical equipment and medical instruments, were evaluated during the full facility survey and measured at baseline and 24 months post-intervention.

Caseload data were collected using a facility survey concurrently collected every 6 months with patient satisfaction and vignette data, which are described below. Caseload scores were calculated using these data: number of outpatients, number of inpatients, number of physicians, and number of days in a month the hospital is in operation. Caseload scores of 1.0 are based on a minimum of 10 patients visited per 8 hours worked. For example, a physician who treats five patients within an 8-hour work period is given a caseload score of 0.5, which is the threshold for passing.

Clinical vignettes

To measure quality of clinical care services (process of care), physicians were administered clinical vignettes, which are open-ended paper cases that simulate and measure actual clinical practice, every 6 months. Vignettes and vignette validation, described elsewhere in detail, represent advances on traditional evaluations of clinical practice quality such as chart abstraction or direct observation (Dresselhaus *et al*. 2000; Peabody *et al*. 2000; Peabody *et al*. 2004a). Vignettes have also been shown to be effective in various international settings, while paediatric vignettes have specifically been shown to be effective in Europe, North America and Asia (Peabody *et al*. 2004b; Peabody *et al*. 2005).

Eligibility for participation was determined by the following criteria: physicians in good standing with licensure and review boards, those who were accredited by the national insurance corporation, and those who work at a PHIC accredited facility. The vignettes were administered to three randomly selected public providers in each hospital every 6 months. Each doctor selected to take the vignettes was required to complete vignettes for three target conditions: pneumonia, diarrhoea, and a common dermatological condition. To complete each vignette, the physicians were asked to take a history, perform a physical examination, select diagnostic tests, make a diagnosis, and provide appropriate treatment. To minimize gaming, we randomly determined the specific vignette sequence for each physician and no physician repeated a vignette in this study.

Once completed, we assigned vignettes randomly to two trained abstractors to score based upon a pre-established scoring system. The explicit scoring criteria are based on WHO clinical practice guidelines and evidence-based criteria, which are used in the Philippines through the Integrated Management of Childhood Illnesses (IMCI). Any scoring discrepancies between scorers were reconciled by a third scorer. An average vignette score for each physician was then calculated. Facility vignette scores were calculated by taking an average of the three selected physicians' vignette scores.

Patient exit survey

Patient exit surveys were administered to guardians of patients between the ages of 6 months and 5 years admitted to one of our 30 facilities during Round 1 and Round 2 of data collection, totalling 2989 and 3053 patients, respectively. The patient exit survey is a brief survey that collects demographic information on the patient, objective and subjective health measures, symptoms related to the illness prior to hospitalization, description of the hospital confinement, perceived satisfaction, provider characteristics, and the total cost of seeking care at the time it is obtained. The exit survey was administered by trained medical technologists 1 day prior to or upon discharge of the patient. Mini patient exit surveys, an abridged version of the patient exit survey, were administered to 10 patients per facility on a quarterly basis.

The patient satisfaction component of the exit survey uses the PSQ-18. The PSQ-18 is a short form of the Patient Satisfaction Questionnaire III (PSQ-III) (Marshall and Hays 1994). The PSQ-III is a 50-item questionnaire that covers the seven domains of general satisfaction, technical quality, interpersonal skills, communication, finances, amount of time spent with the provider, and access to care. The PSQ-18 contains a total of 18 questions that cover the seven domains of quality.

Performance measure—Q*

Q* is a ratio expressed in percentage terms and calculated by computing a weighted average of 70% for the vignette scores, 20% for patient satisfaction, and 10% for case load. These weights were determined based on expert opinion and policy preferences regarding the value of physician performance, patient satisfaction, and doctor/patient ratios in relation to quality of care. Alternative specifications derived from a principal component analysis for Q* were performed as well as a sensitivity evaluation of the weights. These showed that results did not differ significantly with different weighting schemes.

Once the elements of Q* data were collected, these data were encoded and calculated by staff working in the QIDS central office. Q* scores, referred to by the national government as issuances, were distributed by PHIC to those facilities that qualified. After review of a cross-national study that showed the average quality scores between countries to be 60.2–62.6% and in consultation with national policymakers, the cut-off for a passing Q* score was set at 65% (Peabody and Liu 2007).

Feedback of vignette and Q* results was multi-faceted. First, individual vignette results were reported to physicians through mailings. Secondly, the chief of the hospital was given the average vignette score of all the participating physicians, as well as the Q* score for their facility. Individual vignette scores were not revealed to the hospital chief. Thirdly, province governors were given the average vignette score and the Q* score of the hospitals within their province. Finally, hospital chiefs, mayors and governors were briefed on study results through presentations every 6 months.

Data analysis

For all variables, basic descriptive statistics were calculated to check distributions and find outliers and out-of-range data. A multivariate model was used to generate difference-in-difference estimates of Q* over time and across intervention and control groups. Multivariate models for the structural variables were also run to generate difference-in-difference estimates. We report on the change in variables (Q* and structural measures) over time for the intervention and control sites relative to the baseline average score. The same analysis is done for the structural variables. The model used for the Q score can be summarized by the following equation:

$$Qscore_{it} = \alpha + \gamma B_i + \Sigma_t \beta_t B_i P_t + \Sigma_t \theta_t C_i P_t + \varepsilon_i$$

where $Qscore_{it}$ is the Q* score of the i^{th} district hospital in semester t; B_i is a dummy variable indicating where B intervention was introduced; C_i is a dummy variable for QIDS control sites. B_iP_t and C_iP_t are interaction terms between time in 6-month periods (P_t) and intervention type.

Cost of implementation

We calculated the cost of implementing Q* by tallying all associated variable and fixed costs. These costs were for: administering vignettes, physician surveys and facility surveys; entering and encoding the collected data; analysing the data; disseminating the results to physicians and facilities; and coordinating and facilitating all of these steps. Costs were first determined per survey and then aggregated to determine the total cost per facility per every 6-month period of administration.

Results

Implementation of Q*

We found that the Q* measure was readily introduced into our 30 different study sites distributed over a geographically wide and culturally disparate area. We observed this occurred because data collection did not require a great deal of marginal resources. Data on Q* component indicators are collected through three straightforward mechanisms. The caseload data are obtained from existing facility log books; patient satisfaction, measured in the patient exit survey, is collected by hospital staff and supervised by regional health authorities to ensure timelines and consistency; and the vignettes are administered to physicians by trained regional staff. Collecting data on any of these components, therefore, did not require highly skilled personnel and could be done by existing staff.

While the Q* measurement was entirely practicable and doable, a major key to its sustainability as a measure of quality in this study is its integration into the existing health system. Early on, the introduction and application of the Q* metric owed its success to the collaboration between QIDS investigators and stakeholders. Later, the transparency of the measurement and issuances based on Q* largely facilitated stakeholder buy-in. Provincial governors, hospital directors and physicians were notified of the scores through a regularized distribution scheme. These announcements became anticipated and are used for internal monitoring and in crafting ways of improving performance. At the national level, regular meetings between QIDS/PHIC and the government or hospital staff–while challenging to regularize–provide feedback and explain performance

Table 1 Means and standard deviations of service and structural measures in all facilities ($N = 20$)

Variable	Round 1 (baseline)		Round 2	
	Mean	SD	Mean	SD
Q*	62.4%	0.1	66.4%	0.1
No. of doctors	11.0	5.2	5.9	5.2
No. of prescriptions filled in the last week	339.8	332.2	664.8	776.2
No. of functioning items of equipment (out of 21) (Microscope, Echocardiogram, Centrifuge, Defibrillator, X-ray, Anesthesia machine, Ultrasound, Operating room table, Adult ventilator, Operating room lamp, Child ventilator, Cautery machine, Pulse Oximeter, Casting equipment, Cardiac Monitor, Oxygen delivery, Incubator, Nebulizer, Electrocardiogram, Backup generator, Warming bed for newborns)	15.3	3.6	12.3	5.7
No. of functioning instruments (out of 8) (Sterilizer, Otoscope, Baby scale, Resuscitation equipment, IV tubing, Suturing sets, Regular stethoscope, Sterile disposable latex gloves)	6.2	1.1	4.7	2.7
No. of laboratory tests available (out of 13) (Urinalysis, Serum creatinine test, Fecalysis, Electrolytes, CBC, VDRL or RPR test, Blood typing, Liver function tests, Gram stain, Hepatitis B test, TB sputum stain, Bacterial culture, Serum glucose levels)	9.3	1.8	8.8	1.6
Number of supplies available (out of 12) (Antiseptics, Gram stain, Bandages, Acid fast, Oxygen tank, Pregnancy test strips, Suturing materials, Urine strip, IV tubes, VDRL serology, Gloves, Test for occult blood in stool)	10.4	1.1	9.9	1.8

results. Lastly, Q* measurements are tied to bonus payments to the hospitals, giving the metric meaning and consequence. This works by indicating that hospitals achieving a minimum Q* have received a passing score and bonuses that are distributed to hospital providers. Cut-offs are deliberated regularly so that hurdle rates can be routinely increased as overall system performance improves. The measurement, cut-off determination and issuance cycle for the bonuses is done on a strict quarterly basis to maintain consistency in the procedures.

Table 1 summarizes all the measures of quality collected in hospital facilities, including Q*. The table shows the average scores of Q* in Round 2 and at baseline in Round 1, which was conducted 2 years earlier, before the intervention was put in place. Structural variables, also routinely collected at the hospital level, are listed with Q* in the table.

Sensitivity of Q* to policy change

As expected, Q* scores were found to be similar between intervention and control sites at baseline (Table 2). Initially, at the 6-month and then the 12-month assessments, no significant changes were seen in either site ($P > 0.05$). However, after 18 and 24 months, in the fourth and fifth rounds of Q* determination, the intervention hospital sites had significantly higher Q* scores compared with baseline ($P = 0.04$, $P = 0.03$). During the same time period, Q* scores for control sites increased, but only slightly and the changes were not significant ($P > 0.05$).

Sensitivity of structural measures to policy change

To determine the responsiveness of the structural measures captured in the facility survey (namely number of doctors, number of functioning equipment items, number of functioning instruments, number of prescriptions filled, number of lab tests, and supplies available over time), we first looked at the change in the structural measures over time. Across the four assessments, compared with the baseline the number of lab tests, supplies and prescriptions filled (except for controls at the last assessment period) did not change significantly for either intervention or control sites ($P > 0.05$) (Table 3). Number of doctors declined in both intervention and control sites and number of functioning equipment items and instruments declined only in control sites between baseline and Round 2. None of the variables changed significantly at the $P < 0.05$ level in the intervention sites relative to the control sites in difference-in-difference models (not shown in table).

Table 2 Q* scores of intervention and control sites

	Intervention sites		Control sites	
	Q* (%)	*P*-value[a]	Q* (%)	*P*-value[a]
Baseline	62.20	0.90[b]	62.60	
Post-intervention				
+6 months	63.30	0.73	60.80	0.58
+12 months	67.80	0.09	63.80	0.71
+18 months	68.78	0.04	64.48	0.56
+24 months	69.78	0.03	63.77	0.72

[a]Baseline vs. post-intervention.
[b]Intervention vs. control at baseline.

Q* versus structural measures

We compared the variation in our quality measure, Q*, with the variation of structural measures in our intervention sites, as these were the sites in which policy reforms were specifically implemented to improve quality of care. As seen in Figure 1, Q* improved over time in the intervention sites but, by contrast, the structural measures tended to decline over time, although these were not statistically significant from baseline (Table 3). The constancy and the lack of any clear trend for the structural measures suggest their lack of sensitivity to policy changes.

Sensitivity of Q* to weighting specifications

Lastly, although weighting was based on expert input and policy preference, we were interested to see if different

Table 3 Structural measures for intervention and control sites over time

	Intervention sites		Control sites	
Structural measures	Mean	*P*-value[a]	Mean	*P*-value[a]
No. of doctors				
Baseline	10.50	0.84[b]	10.80	
+6 months	10.50	0.99	10.80	0.99
+12 months	7.80	0.05	7.00	0.01
+18 months	8.10	0.08	6.90	0.00
+24 months	6.87	0.01	5.10	0.00
No. of prescriptions filled in the last week				
Baseline	378.00	0.71[b]	301.50	
+6 months	291.80	0.68	139.90	0.44
+12 months	225.30	0.47	311.60	0.96
+18 months	335.10	0.84	409.50	0.61
+24 months	460.13	0.71	828.60	0.01
No. of lab tests available (out of 13)				
Baseline	8.90	0.51[b]	9.80	
+6 months	8.60	0.82	9.10	0.61
+12 months	8.10	0.54	8.70	0.42
+18 months	8.60	0.82	8.50	0.34
+24 months	8.50	0.78	9.00	0.56
No. of supplies available (out of 12)				
Baseline	10.40	0.98[b]	10.40	
+6 months	10.00	0.78	10.00	0.78
+12 months	8.70	0.22	10.40	0.99
+18 months	9.90	0.73	9.50	0.52
+24 months	9.88	0.73	9.90	0.73
No. of functioning items of equipment (out of 21)[c]				
Baseline	6.20	0.97[b]	6.20	
+24 months	5.75	0.70	3.80	<0.001
No. of functioning instruments (out of 8)[c]				
Baseline	15.50	0.82[b]	15.10	
+24 months	13.25	0.21	11.60	<0.001

[a]Baseline vs. post-intervention.

[b]Intervention vs. control at baseline.

[c]Collected using the full facility survey at baseline and during Round 2 (24 months).

specifications would change our results. For sensitivity assessment we used four additional specifications: (1) equal weights, (2) 50% vignettes, 25% satisfaction and 25% case load; (3) 45% vignettes, 10% satisfaction and 45% case load; and (4) a specification utilizing an index derived from principal component analysis. Our findings, namely that improvements in Q* rose differentially after 18 and 24 months, were robust across the four models.

Cost of implementation

Table 4 shows the costs of generating Q*, which include costs associated with rostering, interview with vignette administration, scoring of vignettes, encoding, dissemination (issuances, mailings and briefings of hospital chiefs, governors and mayors) and staff costs including analysis. For all three instruments per facility, costs totalled US$2133.26, which translates into a cost of US$304.84 per assessment. Fixed costs for the development of the vignettes, encoding software, and determination of the sample frame equalled US$21 093.82. These one time costs are eliminated or reduced when existing materials are utilized.

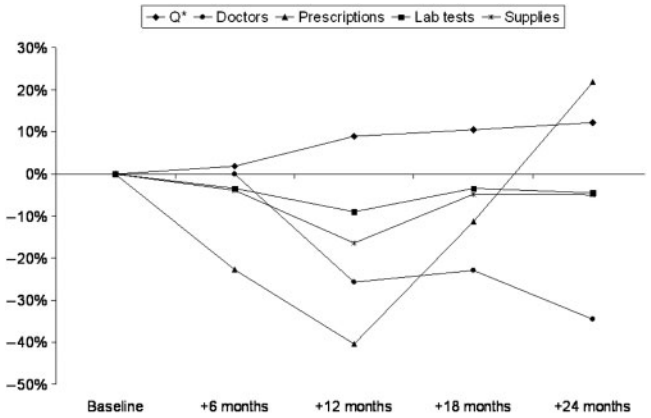

Figure 1 Change in structural measures versus Q* score at each successive semester compared to baseline in Intervention sites

Discussion

We report on the development and implementation of a new metric, Q*, that combines measures of caseload, physician aptitude and patient satisfaction, to detect changes in quality of care in facilities operating within a large national health care system. Q* as a metric reflects the three basic elements of quality of care: structure, process and outcome (Donabedian 1988). We used Q* successfully to serially evaluate system performance over five 6-month periods between December 2004 and December 2006 in 30 participating hospitals covering a large area of the Philippines. Operationally, the Q* system includes data gathering, measurement, issuance, bonus payments and feedback. Q* has been sustainable and acceptable to stakeholders owing to its relative simplicity, transparency and relevancy. We believe that Q* is minimally disruptive and that it is readily grafted onto existing routines of hospitals and payers. It was also affordable, with respect to data collection, computation and feedback.

Unlike structural measures, we found that Q* is more responsive to immediate impacts of policy change. The structural measures did not follow a consistent pattern over time and were divergent, in comparison with the consistent upward trend in Q* over time with the introduction of the intervention. The evaluation of structural measures alone would have pointed to no significant change in sites with policy implementation and would have suggested that the policy was ineffective. In fact, Q* effectively measures marginal increases in performance made possible within significant capacity constraints defined by suboptimal staffing, supplies and equipment. This sensitivity is especially important in a developing country setting where inefficiencies in the delivery of quality care continue.

Table 4 Costs of generating Q*, in US dollars (2005)

Cost per site per round of assessment[a]	Vignettes Sub total ($)	Facility Sub total ($)	Patient exits Sub total ($)
Rostering	18.66		
Vignette administration and physician survey	63.99	6.22	124.50
Scoring of vignettes	32.13		
Encoding	8.01	0.89	
Dissemination	17.79		
Central staff including analysis	16.86	0.85	14.94
Total cost	304.84		

[a]A total of 30 facilities and 7 rounds of data collected.

While structural measures have characteristically been used to assess quality of care received in a hospital, studies have questioned the use of these measures as indicators for quality of care (Peabody *et al*. 1994; Peabody *et al*. 1998; Barber and Gertler 2002). These studies assessing quality of care have typically categorized facility characteristics into four main categories: the physical condition of the clinic building; the availability of basic equipment, sophisticated equipment, supplies and drugs; the staffing level of doctors and nurses; and the availability of laboratory tests (Peabody *et al*. 1994; Peabody *et al*. 1998). Similarly, licensing and accreditation systems require hospitals to meet specific structural standards, such as staffing, bed availability, services available and equipment (Griffith *et al*. 2002). Structural measures, however, are too distant to the interface between patient and provider and do not address whether the inputs are used properly to produce better health (Dresselhaus *et al*. 2000). The assumption that with good equipment and instrumentalities, good care will follow is not necessarily correct and was not seen in our results. By contrast, we found that quality of clinical care could improve even in the absence of new resources, better staffing, increased supplies or investments in equipment. Many measures in this analysis, such as laboratory tests offered and supplies available, did not change over time, even as policy interventions were being implemented. Moreover, neither a clear nor a predictable pattern was seen between changes in structural measures and policy change, revealing a lack of sensitivity to these determinants. While a connection between structural measures and quality of care cannot be entirely dismissed, perhaps the strength of the connection needs to be questioned. Q*, however, was responsive to change and, unlike structural measures, was effective in tracking the impacts of health policy on the quality of care.

In general, hospitals strive to use the least-cost mix of inputs to provide a certain level of quality. If at any point they are cost-minimizing and are fully utilizing all available structural inputs, then any improvement in quality can only be had with an increase in inputs. Put differently, such hospitals will necessarily have to increase resources to effect any further improvement in quality. This is not the scenario implied by our results, which show that while quality, as measured by Q*, increased continuously over the 24-month period, input or structural indicators of quality remained the same. As cost minimization is an expected behaviour of hospitals, our findings reflect the existence of excess capacity to provide quality or that quality improvements are being produced with the existence of excess resources. Accordingly, even if the amount of inputs were reduced or slack eliminated, a hospital could continue to increase quality levels. A second implication is that other interventions such as financial incentives or a system of quality monitoring would be more effective than additional investments in equipment, technology or supplies.

There are some limitations to our study that warrant discussion. The first is around measurement: while care was taken to include a thorough assessment of structural measures, there may have been some for which we could not observe changes over time. Further, while we observe that changes in Q* reflect changes in practice, changes in practice will have to be correlated to change in health outcomes. In terms of vignettes, it is possible that vignette performance in the post-intervention group may have changed disproportionately more (or less) than in the post-intervention control group. Ultimately, to answer this question, future studies will have to either see if there is a divergence of vignette scores versus the gold standard evaluations—something that will be nearly impossible to evaluate logistically—or more practically compare the difference in health outcomes between the two groups.

The quality of health care has begun to receive increased attention because of the intuitive link between it and subsequent health outcomes for patients. In the last 30 years, research has demonstrated that quality can be measured, varies enormously, is affected more by where you go than who you are, and it can be improved, but this is difficult and painful, and, in general, has not been successfully accomplished (Brook *et al*. 2000). Brook and colleagues attribute the last deficiency to the lack of a government policy to support the development of a set of quality assessment tools, in any country in the world. While comprehensive monitoring and evaluation of system performance can assess quality and has the potential to inform and improve health systems, there is always the concern that it requires substantial personnel and monetary resources. Q*, on the other hand, is an efficient quality assessment tool that we have found to be a comprehensive and effective measurement of quality, as well as one that is achievable in a resource-poor setting.

Acknowledgements

The authors would like to acknowledge the U.S. National Institutes of Child Health and Development (NICHD R-01 HD042117) for the funding that supported this work.

The study was reviewed and approved by the university Institutional Review Boards at the University of California, San Francisco and the University of the Philippines. Written, free and full informed consent was obtained by all study participants.

There are no competing interests to declare.

References

Almeida C, Braveman P, Gold MR *et al.* 2001. Methodological concerns and recommendations on policy consequence of the World Health Report 2000. *The Lancet* **357**: 1692–7.

Arah OA, Westert GP, Hurst J, Klazinga NS. 2006. A conceptual framework for the OECD Health Care Quality Indicators Project. *International Journal of Quality in Health Care* **18**: 5–13.

Barber SL, Gertler PJ. 2002. Child health and the quality of medical care. Working paper. University of California, Berkeley. Online at: http://faculty.haas.berkeley.edu/gertler/working_papers/02.28.02_childheight.pdf, accessed 23 March 2007.

Bouchet B, Francisco M, Ovretveit J. 2002. The Zambia quality assurance program: successes and challenges. *International Journal of Quality in Health Care* **14** (Suppl. 1): 89–95.

Braveman P, Starfield B, Geiger HJ. 2001. World Health Report 2000: how it removes equity from the agenda for public health monitoring and policy. *British Medical Journal* **323**: 678–81.

Brook RH, McGlynn EA, Shekelle PG. 2000. Defining and measuring quality of care: a perspective from US researchers. *International Journal for Quality in Health Care* **12**: 281–95.

Donabedian A. 1988. The quality of care: How can it be assessed? *Journal of the American Medical Association* **260**: 1743–48.

Donabedian A. 2005. Evaluating the quality of medical care. *The Milbank Quarterly* **83**: 691–729.

Dresselhaus TR, Peabody JW, Lee M, Glassman P, Luck J. 2000. Measuring compliance with preventive care guidelines: A comparison of standardized patients, clinical vignettes and the medical record. *Journal of General Internal Medicine* **15**: 782–8.

Griffith JR, Knutzen SR, Alexander JA. 2002. Structural versus outcomes measures in hospitals: a comparison of Joint Commission and Medicare outcomes scores in hospitals. *Quality and Management in Health Care* **10**: 29–38.

Mainz J. 2003. Defining and classifying clinical indicators for quality improvement. *International Journal for Quality in Health Care* **15**: 523–30.

Marshall GN, Hays RD. 1994. The Patient Satisfaction Questionnaire Short-Form (PSQ-18). Document Number: P-7865. Santa Monica, CA: RAND.

McNamara P. 2005. Quality-based payment: six case examples. *International Journal of Quality in Health Care* **17**: 357–62.

McNamara P. 2006. Purchaser strategies to influence quality of care: from rhetoric to global applications. *Quality and Safety in Health Care* **15**: 171–3.

Murray CJ, Frenk J. 2006. A WHO Framework for Health System Performance Assessment. Global Programme on Evidence, Working Paper No.6. Geneva: World Health Organization.

Ovretveit J. 2004. Formulating a health quality improvement strategy for a developing country. *International Journal of Health Care Quality Assurance* **17**: 368–76.

Peabody JW, Liu A. 2007. A cross-national comparison of the quality of clinical care using vignettes. *Health Policy and Planning* **22**: 294–302.

Peabody JW, Rahman O, Fox K, Gertler P. 1994. Quality of care in public and private primary health care facilities: Structural comparisons in Jamaica. *Bulletin of the Pan American Health Organization* **28**: 122–41.

Peabody JW, Gertler PJ, Leibowitz A. 1998. The policy implications of better structure and process on birth outcomes in Jamaica. *Health Policy* **43**: 1–13.

Peabody JW, Luck J, Glassman P, Dresselhaus TR, Lee M. 2000. Comparison of vignettes, standardized patients, and chart abstraction: a prospective validation study of 3 methods for measuring quality. *Journal of the American Medical Association* **283**: 1715–22.

Peabody JW, Luck J, Glassman P *et al.* 2004a. Measuring the quality of physician practice by using clinical vignettes: a prospective validation study. *Annals of Internal Medicine* **141**: 771–80.

Peabody JW, Tozija F, Munoz JA, Nordyke RJ, Luck J. 2004b. Using vignettes to compare the quality of care variation in economically divergent countries. *Journal of Health Services Research* **39**: 1951–70.

Peabody JW, Taguiwalo MM, Robalino DA, Frenk J. 2006. Improving the quality of care in developing countries. In: Jamison DT, Breman JG, Measham AR *et al.* (eds). *Disease Control Priorities in Developing Countries*. London: Oxford University Press, pp. 1293–307.

Shimkhada R, Peabody JW, Quimbo SA, Solon O. 2008. The Quality Improvement Demonstration Study: An example of evidence-based policy-making in practice. *Health Research Policy and Systems* **6**: 5.

Soeters R, Habineza C, Peerenboom PB. 2006. Performance-based financing and changing the district health system: experience from Rwanda. *Bulletin of the World Health Organization* **84**: 884–9.

World Health Organization. 2000. *The World Health Report 2000. Health Systems: Improving Performance*. Geneva: World Health Organization. Online at: http://www.who.int/whr/2000/en/whr00_en.pdf, accessed 23 March 2007.