International Journal for Quality in Health Care 2011; pp. 1–7 10.1093/intqhc/mzr057

Validity of self-assessment in a quality improvement collaborative in Ecuador

JORGE HERMIDA[1], EDWARD I. BROUGHTON[2,3] AND LYNNE MILLER FRANCO[2,4]

[1]University Research Co, LLC, Quito, Ecuador, [2]UniversityResearchCo,LLC,ChevyChase,MD,USA, [3]Johns Hopkins School of Public Health, Baltimore, MD, USA, and [4]EnCompass LLC, Rockville, MD, USA

Address reprint requests to: Jorge Hermida, University Research Co, LLC/HCI, Avenida ColonE8-57 y Diego Almagro, Edificio El Cisne, Sto Piso Oficina C-D, Quito, Ecuador. Tel: +11-593-2-2-222-120; Fax: +11-593-2-2-222-119; E-mail: jhermida@urc-chs.com

Accepted for publication 18 July 2011

Abstract

Objectives. Health care quality improvement (QI) efforts commonly use self-assessment to measure compliance with quality standards. This study investigates the validity of self-assessment of quality indicators.

Design. Cross sectional.

Setting. A maternal and newborn care improvement collaborative intervention conducted in health facilities in Ecuador in 2005.

Participants. Four external evaluators were trained in abstracting medical records to calculate six indicators reflecting compliance with treatment standards.

Interventions. About 30 medical records per month were examined at 12 participating health facilities for a total of 1875 records. The same records had already been reviewed by QI teams at these facilities (self-assessment).

Main Outcome Measures. Overall compliance, agreement (using the Kappa statistic), sensitivity and specificity were analyzed. We also examined patterns of disagreement and the effect of facility characteristics on levels of agreement.

Results. External evaluators reported compliance of 69–90%, while self-assessors reported 71–92%, with raw agreement of 71–95% and Kappa statistics ranging from fair to almost perfect agreement. Considering external evaluators as the gold standard, sensitivity of self-assessment ranged from 90 to 99% and specificity from 48 to 86%. Simpler indicators had fewer disagreements. When disagreements occurred between self-assessment and external valuators, the former tended to report more positive findings in five of six indicators, but this tendency was not of a magnitude to change program actions. Team leadership, understanding of the tools and facility size had no overall impact on the level of agreement.

Conclusions. When compared with external evaluation (gold standard), self-assessment was found to be sufficiently valid for tracking QI team performance. Sensitivity was generally higher than specificity. Simplifying indicators may improve validity.

Keywords: quality measurement, audit, quality indicators, statistical methods, reproductive health, hospital care, general medicine

Introduction

Quality improvement collaboratives (QICs) are interventions that use a shared learning approach among a large number of QI teams who work together on the same problem area to rapidly achieve significant improvements in processes, quality and efficiency of services. Most literature on QICs in health care relies on QI teams performing self-assessments of compliance with quality standards [1]. This is often the most efficient method of data collection for performance indicators and is therefore frequently used in resource-constrained settings [2]. Some have found health provider self-assessment to be effective in improving performance in QI interventions where higher level supervision is unavailable [3]. Information from such assessment is crucial to design QIC interventions, identify performance gaps that require attention and allow the QI team to monitor its progress in improving health-care delivery processes [4]. It is therefore crucial that these data be a valid representation of performance.

However, some authors have questioned the overall validity of self-assessment in QI interventions [1]. Mittman [5] states that self-reported data are likely to be biased in favor of positive findings, although he offers no evidence to support his hypothesis. Outside the context of continuous quality improvement (CQI) programs, studies examining the validity of self-assessment have been mixed, with a greater

number demonstrating bias toward self-enhancement rather than self-diminishment [6]. In a comparison of medical records, patient surveys and provider surveys regarding tobacco counseling during primary care visits, Conroy *et al.* [7] found little agreement among the three with provider respondents consistently stating they fulfilled the requirements more frequently than indicated by the other two measures. A study of dental health providers found variable agreement between self-reports of procedures used and direct observations and there was consistent over-estimation of performance by the providers [8]. There is little peer-reviewed literature that specifically examines the validity of self-assessment of performance of QI interventions [9].

QIC activities conducted by 49 Ministry of Health hospitals and 27 health centers in Ecuador as part of an essential obstetric care [10] collaborative offered a unique opportunity to examine the validity of self-assessment. These 76 facilities, located throughout half of Ecuador's 22 provinces use monthly review of clinical records by QI team members to monitor and continuously improve the quality of maternal and child services they provide. Ministry of health staff located at the provincial offices act as coaches (QIC Facilitators) to provide periodic supervision and support visits to facilities [11].

This study examines the reliability of structured self-assessment by QI teams of performance, measured as compliance with specific standards of care, compared with assessment by external evaluators in a sample of teams participating in the collaborative. Second, it evaluates whether there is a tendency toward reporting performance enhancement or diminishment in the disagreements between self-assessment and external evaluators. Third, it determines the association between the degree of validity of self-assessment data and factors such as facility size, the level of leadership and support given by the QIC facilitators and the degree of understanding reported by QI teams of the instructions and tools for collecting information for the indicators.

Methods

Data collection

This study focused on comparison of self-assessment and external evaluation of performance in a purposive sample of 12 of the 49 hospitals from the six provinces participating in the collaborative. All six participating provincial hospitals were included. From each of the same six provinces as these hospitals, we also selected the participating county hospital with the highest number of births in the previous year. These were selected to increase the likelihood that at least 30 births occurred in the facility for the month of observation.

All QI teams used the same methodology for conducting a review of their performance—taking a sample of clinical records for services provided during the month, and applying a structured clinical record abstraction sheet to record whether specific standards had been complied with. QI teams reviewed a sample of 30 clinical records each month based on a systematic sample with a random start (i.e. all clinic records available for review were divided by 30 to obtain the sampling interval). In some low-volume facilities, the sample of 30 included all (or almost all) of the patients seen in the month. A total of 1875 patient records were examined. The QI teams did not know the records would be reexamined as part of this study.

QI teams monitored six indicators, each related to a specific standard (Table 1). Each standard had objective, specific verifiable criteria in written format [12] that QI teams used to determine whether that standard had been complied with, and every QI team was trained to apply these criteria. Each indicator was expressed, discussed and reported as a percentage of encounters for that type of service for which all criteria for that standard were complied with. If one single criterion was not met, the record was considered non-compliant for the indicator.

Subsequently, four experienced consultants independently visited the 12 selected hospitals and re-assessed the same clinical records the QI teams had assessed, using the same standards and criteria. The consultant determined whether the clinical record indicated that all criteria had been adhered to for each specific standard. The four consultants were trained in the use of the specific criteria, and their inter-observer agreement was tested until their agreement level was higher than 90% for standards being measured.

Additionally, three characteristics of the facility or the collaborative facilitator performance were recorded for all 12 sites: (i) leadership—a score from 1 to 10, given by QI teams on the quality of leadership provided by their coach; (ii) the self-reported degree of understanding of the tools used in the QI self-assessment process; and (iii) the volume of deliveries seen in the facility.

Data analysis

The level of agreement was measured using the Kappa statistic, which compares the observed concordance found between the results obtained from self-assessment and from external evaluators and the concordance expected due to chance alone. Point estimates for the Kappa statistics were classified according to the descriptions used by Viera [13] and Garrett.

Where there was disagreement between self-assessment and external evaluator findings, we tested whether they occurred by chance or if there was evidence of bias toward self-enhancement in QI team performance, using Fisher's exact test.

Sensitivity (the probability of self-assessment finding non-compliance with an indicator when the external evaluator reported non-compliance) and specificity (the probability of self-assessment finding compliance with an indicator when the external evaluator reported compliance) were determined using two-by-two contingency tables where the results reported by external evaluators were considered the 'gold standard' and non-compliance with a standard was considered a 'positive' test result. In cases where either one or both the external evaluator and self-assessment findings were 'not applicable', the observations were excluded from sensitivity and specificity calculations. Results for categories where there were multiple indicators were averaged.

Table 1 Indicators and their specific criteria

Indicator	Specific criteria
Proportion of prenatal care visits for which all criteria were adhered to	1. Was personal and obstetric history taken? 2. Was mother's height recorded? 3. Was breast examination completed? 4. Was number of weeks of amenorrhea recorded? 5. Was body weight measured? 6. Was blood pressure measured? 7. Was uterine height measured? 8. Was fetal heart rate recorded? 9. Was fetal movement recorded? 10. Was edema recorded? 11. Was genital bleeding evaluated? 12. Was nutritional assessment completed? 13. Was tetanus vaccination ordered or registered? 14. Were prenatal laboratory tests ordered? 15. Was dental examination ordered?
Proportion of deliveries in which a partograph was used	1. Was a partograph used during delivery?
Proportion of deliveries in which all criteria were adhered to	1. Were curves correctly plotted on partograph? 2. Was BP monitored and recorded? 3. Were frequency and duration of contractions monitored? 4. Was fetal HR monitored/recorded?
Proportion of deliveries in which oxytocin was applied correctly in third stage of labor	1. Was oxytocin given within 1 min of birth?
Proportion of deliveries in which all criteria for *post partum* care were adhered to	1. Was time recorded? 2. Was temperature taken/recorded? 3. Was pulse taken and recorded? 4. Was BP taken and recorded? 5. Was uterine involution assessed/recorded? 6. Were characteristics of vaginal discharge recorded?
Proportion of live births for which all immediate newborn care criteria were adhered to	1. Was sex of infant registered? 2. Was infant's birthweight recorded? 3. Was infant's height recorded in cm? 4. Was infant's head circumference recorded in cm? 5. Was APGAR score recorded 6. Was there a record of resuscitation performed? 7. Was a physical examination reported? 8. Was there a record of infant being kept with mother in same room? 9. Was the start of breastfeeding recorded?

We also tested to determine whether or not disagreements were biased towards self-enhancement by the QI teams. For each indicator, we compared the proportion of times the QI teams mistakenly reported compliance with the proportion of times they mistakenly reported non-compliance. If no bias was present, these proportions would be the same.

Next, the total number of disagreements was reported across criteria for each of the six standards listed in Table 1.

Finally, we used linear regression to analyze the relationship between the number of disagreements and the QI teams' understanding of the indicators, facilitator leadership and support, and the volume of deliveries occurring at the facility.

Results

Compliance level according to self-assessment and external evaluation

According to both self-assessment and assessment by external evaluation, the two indicators with highest levels of compliance were prenatal care and immediate newborn care, while the indicator with the lowest compliance was use of oxytocin as part of active management of the third stage of labor (AMTSL). The external evaluators consistently found lower compliance than the QI teams' self-assessment, although for many indicators, the differences were small;

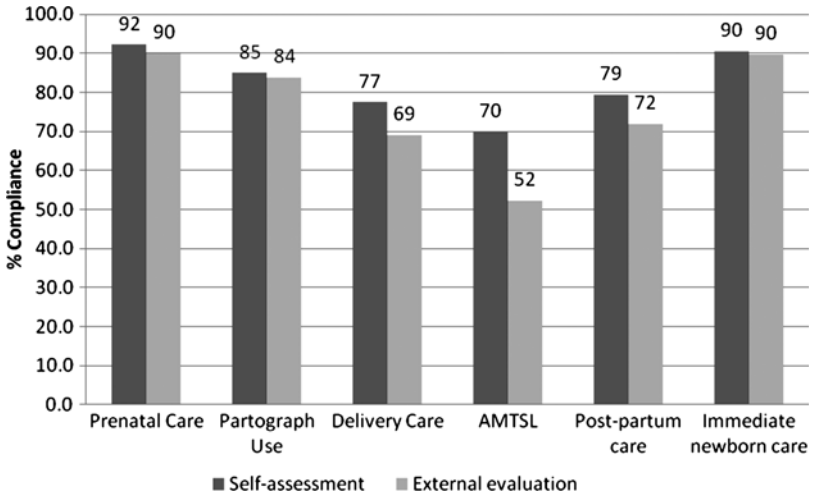

Figure 1 Reported compliance by self-assessment and external evaluation.

Table 2 Agreement, Kappa statistics, sensitivity and specificity on compliance evaluation by self-assessment and external evaluators

	% Agreement	Kappa	Mean sensitivity	Mean specificity
Prenatal care	90	0.44	97	48
Partograph Use	86	0.51	96	83
Delivery care	71	0.36	92	54
AMTSL	71	0.42	90	54
Post partum care	89	0.71	99	71
Immediate newborn care	95	0.82	98	86

differences of closer to 10% were seen for delivery care and *post partum* care and 17.6% for the AMTSL indicator (Fig. 1).

Kappa, sensitivity and specificity

The percent agreement between the external evaluators and self-assessment was highest for immediate newborn care and lowest for AMTSL and delivery care. When the proportion of agreement expected due to chance was taken into consideration with the Kappa statistics, *post partum* and newborn care were rated very highly while delivery care was the lowest, rating at 'fair' (Table 2). Mean sensitivity was high for all indicators and close to 100% for partograph use, prenatal, *post partum* and immediate newborn care. Mean specificity was generally low, ranging between 48% for prenatal care and 86% for immediate newborn care.

Number of disagreements

Each indicator contained a set of specific criteria to measure attainment of the standard. Examination of the proportion of agreement indicated that the lowest level of agreement was for the prenatal care indicator, which was as expected given there was the largest number of criteria (15) that make up this indicator, and therefore more opportunities for errors. Forty-nine percent of charts examined for the delivery care indicator reliability had no disagreements, while 12% had three disagreements out of a possible four. For both immediate *post partum* and neonatal care indicators, >80% of charts reviewed showed complete agreement (Table 3).

Pattern of disagreements

For five of the six indicators, there were statistically significant differences between the observed proportions of disagreements—where the QI teams reported compliance with standards but the external evaluators did not—and disagreements expected due to chance ($P < 0.01$ for all five). Only for immediate newborn care did it appear that the pattern of disagreement was due to chance (observed = 0.50, expected = 0.50, $P = 0.94$).

Factors associated with disagreements

Scores assigned by the QI teams on their facilitator's leadership, reported levels of understanding of instructions and tools and facility size measured by the number of births they performed annually were not associated with the proportion of disagreements in the indicators in any discernable pattern (Table 4).

Discussion

This study demonstrated high general agreement across the six essential obstetric and newborn care indicators between

Table 3 Number and percent disagreements between self-assessment and external evaluation for individual criteria that comprise each standard

Disagreements	Prenatal care		Partograph use		Delivery care		AMTSL		*Post partum* care		Immediate newborn care	
	Frequency	%	Frequency	%	Frequency	%	Frequency	%	Frequency	%	Frequency	%
0	183	44.7	625	86.45	296	49.3	422	71.4	590	84.9	568	82.8
1	68	16.6	98	13.55	180	30.0	169	28.6	26	3.7	82	12.0
2	41	10.0			17	2.8			10	1.4	16	2.3
3	57	13.9			37	6.2			5	0.7	9	1.3
4	31	7.6			71	11.8			3	0.4	3	0.4
5	8	2.0							4	0.6	2	0.3
6	7	1.7							57	8.2	2	0.3
7	7	1.7									1	0.2
8	3	0.7									2	0.3
9	1	0.2									1	0.2
10	3	0.7										
Total	409	100	723	100	601	100	591	100	695	100	686	100

Table 4 Variables affecting the odds of one or more disagreements in indicators

	Variable	Odds	*P*-value	
Prenatal care	Leadership by facilitator	0.94	0.577	
	Understanding of instructions and tools	1.35	0.200	
	Facility with >2000 births	1.67	0.349	
Partograph use	Leadership by facilitator	1.14	0.175	
	Understanding of instructions and tools	2.50	0.001	**
	Facility with >2000 births	2.41	0.208	
Delivery care	Leadership by facilitator	1.08	0.274	
	Understanding of instructions and tools	1.38	0.200	
	Facility with >2000 births	1.58	0.261	
AMTSL	Leadership by facilitator	0.71	<0.001	**
	Understanding of instructions and tools	0.83	0.605	
	Facility with >2000 births	2.62	0.111	
Post partum care	Leadership by facilitator	0.80	0.001	**
	Understanding of instructions and tools	0.87	0.731	
	Facility with >2000 births	2.83	0.123	
Immediate newborn care	Leadership by facilitator	0.94	0.529	
	Understanding of instructions and tools	1.54	0.078	
	Facility with >2000 births	3.66	0.001	**

*Denotes statistical significance, $P < 0.01$.

QI teams and external evaluators using the same methods on the same clinical records.

Kappa statistic calculations, which account for agreement due to chance, revealed levels of agreement of fair or better. Kappa statistics are most revealing when proportion values are not at the extremes (very high or very low). Because levels of compliance with standards were generally high, Kappa calculations sometimes revealed only fair or moderate levels, even though percent agreements were high. Overall, the results indicate a relatively high degree of internal validity of structured self-assessment by QI teams.

Two of the six indicators had higher levels of disagreement between QI team self-assessment and external evaluator results and lower Kappa statistics: AMTSL and delivery care indicators. These two also posed the most difficulty in achieving reliability among the four external evaluators

during their training. For the AMTSL indicator, many references to oxytocin in clinical records did not specify whether it was administered to induce labor or prevent *post partum* hemorrhage, and extra time was needed to examine medical records to understand when it was administered. For delivery care indicator, criteria for determining if the standards reflecting 'correct' partograph use are complex because they are related to variables such as the mother's position during labor, rupture of the amniotic sac, the number of previous deliveries, and other factors. By the end of the training, external evaluators achieved consistently high agreement among themselves for all six indicators, but this experience suggests that the nature of the indicators themselves may have been more important in disagreements between QI teams and external evaluators than errors by the QI teams.

Sensitivity was 90% or above indicating that QI teams are able to correctly identify clinical records in which patient care adhered to standards. However, specificity was consistently much lower, indicating that structured self-assessment is less able to identify non-compliance in clinical records. While lower specificity is expected when compliance levels are high, it is also a cause of some concern, given that identifying non-compliance is supposed to stimulate action by the QI team to rectify any problems occurring in clinical practice. This result suggests that further effort by QI teams to identify deficiencies would be beneficial.

Where disagreements did occur between teams and external evaluators, there was a tendency for QI teams to err on the side of self-enhancement of performance. This concurs with the findings of Conroy *et al.* [7] and, outside the medical field, of Falchikov and Boud [14] and Harris and Schaubroeck [15]. However, others have found a tendency for clinicians score their performance accurately [16, 17] or to underestimate their adherence to standards [18]. Lu and Ma [19] found that incentives for performance may lead to provider misreporting due to self-interest. The degree to which self-enhancement did occur in this study, while statistically significant for some indicators, was not of a large enough magnitude to alter the decision on a potentially needed QI intervention. This absence of practical significance of self-enhancement suggests that while it is of concern and should be rectified, it is not a factor that substantively weakens the validity of self-assessment.

Houston *et al.* [17], in their study on medical residents' accuracy in abstracting their own charts to determine performance compared with performance determined by trained abstractors, found similar results for both groups, with agreement all $>80\%$, Kappa statistics mostly >0.7, sensitivity between 80 and 100% and specificity generally much lower than sensitivity. In a study that compared self-reported provider adherence with tobacco cessation program guidelines to adherence according to electronic medical records and patient surveys, Conroy *et al.* [7] found little agreement between the three assessment methods with the providers reporting higher adherence in four of the five criteria.

Our findings showed that for five of the six indicators, there was evidence to support the hypothesis that when there was disagreement between the findings from self-assessment and external evaluation, the errors favored self-enhancement of QI team performance. It was beyond the scope of this investigation to determine reasons for this pattern. It is possible that QI teams are erring on the side of designating a medical chart compliant with a standard when there is uncertainty, whereas the external evaluators are not. There may possibly be extrinsic rewards for higher compliance levels in the form of praise from CQI facilitators and pride when presenting results at provincial meetings.

It was an unexpected finding that there was no discernible association between facility characteristics and self-assessment validity. In particular, we expected a difference in validity between large and small facilities because the nature of the QI teams and their work is quite different. However, in only one of the six indicators were self-assessments from larger facilities significantly better than those from smaller facilities. We also expected higher validity in teams whose facilitators were perceived as more supportive, but this was not the case.

Limitations

This study examined internal validity—the consistency of the findings of structured self-assessment when measured against the 'gold standard' review by external evaluators. It does not address the question of the construct validity of QI team assessment of clinical records to indicate actual quality of clinical performance. Basing an assessment of quality of care on information from a written chart assumes the chart accurately reflects what services were provided and the manner in which they were delivered. For example, a medical record of a delivery may indicate that oxytocin was administered within one minute of delivery. The method used in this study determines whether a QI team member and an external evaluator agreed that the written record indicates appropriate administration of oxytocin, an important AMTSL indicator component. It does not examine whether the oxytocin truly was given according to guidelines. Study designs that do address this issue—tests of agreement between self-assessment and either observation of service delivery or exit interviews with patients—have significant flaws themselves and it was beyond this study's scope to follow this line of inquiry. However, it is an issue that warrants further investigation.

As mentioned previously, external evaluators, at the beginning of their training, were not always in agreement among themselves in their assessment of identical clinical records, particularly for some indicators. While we addressed this by training the external evaluators until there was consistent unanimity of judgments on compliance in the records, similar in-depth discussion and training were not provided to QI teams.

This research was conducted in a specific setting and the results may not be applicable to self-assessment conducted in different health settings. Similar research in a wider range of settings will yield more generalizable findings.

Policy recommendations

The primary finding is that self-assessment of compliance with quality of care standards reported in clinical records is a

valid way to track performance in this setting. While there was a tendency for QI teams to err towards self-enhancement of their performance, the degree to which it occurs does not adversely influence the results. Assessment of more complicated indicators resulted in a greater number of disagreements between QI teams and the external evaluators. We recommend that the Ministry of Health of Ecuador, which is supporting QI interventions in all of its facilities, continue to use self-assessment for evaluating quality performance. The study showed that the complexity of indicators is important in the validity of self-assessment data. We recommend to those designing QI programs that they use indicators that are simple and explicit to help improve validity of self-assessed measures taken by QI teams in the field.

This research also demonstrates a simple methodology that can be used to assess the validity of self-assessment in QIC programs. The program in Ecuador and similar QI interventions in other countries that rely on self-assessment to monitor provider performance could adopt this method for periodic monitoring of self-assessment validity and address issues as they arise.

Conclusions

The findings support continued use of structured self-assessment for tracking QI team performance and informing changes in their interventions to maximize their positive effect in this setting. It highlights the need for careful selection of indicators for measuring quality of care that can produce reliable, valid results, regardless of who is undertaking the assessment. It also demonstrates a sound, efficient methodology for evaluating the validity of structured self-assessment that can be used in similar settings. Studies of the validity of self-assessment in other countries are required to determine the international generalizability of these findings.

Acknowledgements

We thank the Ecuador Ministry of Health participating hospitals and their QI teams for their support in this study.

Funding

This work was supported by the American people through the United States Agency for International Development (USAID) and its Quality Assurance Project (QAP) [contract number GHN-I-01-00003-00] and Health Care Improvement (HCI) [contract number GPH-C-00-02-00004-00].

References

1. Schouten LM, Hulscher ME, van Everdingen JJ *et al*. Evidence for the impact of quality improvement collaboratives: systematic review. *BMJ* 2008;**336**:1491–4.
2. Franco L, Marquez L, Ethier K *et al*. Results of collaborative improvement: effects on health outcomes and compliance with evidence-based standards in 27 applications in 12 countries. *Collaborative Evaluation Series*. Health Care Improvement Project. ChevyChaseMD:UniversityResearchCo.,LLC(URC),2009.
3. Kelley E, Kelley AG, Simpara CH *et al*. The impact of self-assessment on provider performance in Mali. *Int J Health Plann Manage* 2003;**18**:41–8.
4. Vos L, Duckers ML, Wagner C *et al*. Applying the quality improvement collaborative method to process redesign: a multiple case study. *Implement Sci* 2010;**5**:19.
5. Mittman BS. Creating the evidence base for quality improvement collaboratives. *Ann Intern Med* 2004;**140**:897–901.
6. John OP, Robins RW. Accuracy and bias in self-perception: individual differences in self-enhancement and the role of narcissism. *J Pers Soc Psychol* 1994;**66**:206–19.
7. Conroy MB, Majchrzak NE, Silverman CB *et al*. Measuring provider adherence to tobacco treatment guidelines: a comparison of electronic medical record review, patient survey, and provider survey. *Nicotine Tob Res* 2005;**7**(Suppl. 1):35–43.
8. Demko CA, Victoroff KZ, Wotman S. Concordance of chart and billing data with direct observation in dental practice. *Community Dent Oral Epidemiol* 2008;**36**:466–74.
9. Bose S, Oliveras E, Edson WN. How can self-assessment improve the quality of healthcare? Operations Research Issue Paper. ChevyChase, MD: Quality Assurance (QA) Project and JHPIEGO Corporation, 2001.
10. Hermida J. Scaling up and institutionalizing continuous qulaity improvement in the free maternity and child care program in Ecuador. LACHSR Report No 65. Published for the U.S. Agency for International Development by the Quality Improvement Project, 2005.
11. Hermida J, Robalino ME. Increasing compliance with maternal and child care quality standards in Ecuador. *Int J Qual Health Care* 2002;**14**(Suppl. 1):25–34.
12. The Quality Assurance Project. *Free Maternity Program: Quality Standards and Indicators*. Quito, Ecuador: QAP, Ecuador Ministry of Health, 2003.
13. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. *Fam Med* 2005;**37**:360–3.
14. Falchikov N, Boud D. Student self-assessment in higher education: a meta-analysis. *Rev Educ Res* 1989;**59**:395–430.
15. Harris MM, Schaubroeck J. A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. *Personnel Psychol* 1988;**41**:43–62.
16. Tsugawa Y, Tokuda Y, Ohbu S *et al*. Professionalism mini-evaluation exercise for medical residents in Japan: a pilot study. *Med Educ* 2009;**43**:968–78.
17. Houston TK, Wall TC, Willet LL *et al*. Can residents accurately abstract their own charts? *Acad Med* 2009;**84**:391–5.
18. Braend AM, Gran SF, Frich JC *et al*. Medical students' clinical performance in general practice—triangulating assessments from patients, teachers and students. *Med Teach* 2010;**32**:333–9.
19. Lu M, Ma CT. Consistency in performance evaluation reports and medical records. *J Ment Health Policy Econ* 2002;**5**:141–52.