East African Medical Journal Vol: 93 No. 4 April 2016
INVESTIGATION OF THE SEQUENTIAL VALIDITY OF QUALITY IMPROVEMENT TEAM SELF-ASSESSMENTS IN A HEALTH FACILITY HIV IMPROVEMENT COLLABORATIVE IN TANZANIA
S. N. Kinoti; MBChB, MMED, MPSID, Senior Improvement Advisor, University Research Co LLC; 10753 Gloxinia Drive, Rockville MD, 20852, B. R. Burkhalter, PhD, Senior Research Advisor, University Research Co LLC; 7200 Wisconsin Avenue, Bethesda MD, 20814 (Posthumous), D. Rumisha, MD, MPH; URC Country Director and Chief of Party for Health Care Improvement Project, Skyways Building, 1st Floor, Ohio Street, P. O. Box 71561, Dar es Salaam, Tanzania, E. Hizza, Senior Technical Advisor, MD, MPH, Health Care Improvement Project, Skyways Building, 1st Floor, Ohio Street, P. O. Box 71561, Dar es Salaam, Tanzania, M. Ngonyani, Senior Nutritionist, Health Care Improvement Project, MPH Skyways Building, 1st Floor, Ohio Street, P.O. Box 71561, Dar es Salaam, Tanzania, E. Broughton, PhD, Senior Research DirectorUniversity Research Co LLC; 7200 Wisconsin Avenue, Bethesda MD, 20814 and T. Gondwe, MPH Research Fellow, Community Systems Foundation, University of North Carolina, 250 E. Franklin Street, Chapel Hill, NC 27514

Request for reprints to: S. N. Kinoti; Senior Improvement Advisor, University Research Co LLC; 10753 Gloxinia Drive, Rockville MD, 20852

INVESTIGATION OF THE SEQUENTIAL VALIDITY OF QUALITY IMPROVEMENT TEAM SELF-ASSESSMENTS IN A HEALTH FACILITY HIV IMPROVEMENT COLLABORATIVE IN TANZANIA

S. N. KINOTI, B. R. BURKHALTER, D. RUMISHA, E. HIZZA, M. NGONYANI, E. BROUGHTON and T. GONDWE

ABSTRACT

***Background*: Self-assessment is widely used in the health care improvement collaboratives quality improvement (QI) teams' to assess their own performance. There is mixed evidence on the validity of this approach. This study investigated sequential validity of self-assessments in a QI HIV collaborative in Tanzania.**
***Objectives*: Define the separate self-assessment steps in QI process; determine if the validity of self-assessments improved over time; determine if validity improvement is the same for the different self-assessment activities and determine if validity is the same for the different facilities and type of care.**
***Design*: Prospective semi-quantitative study.**
***Setting*: The study was undertaken over 10 months in nine facilities in Mtwara region of Tanzania following appropriate approvals. Study did not interfere with routine services and processes of continuous quality improvement at the facilities.**
***Subjects*:Trained investigators retrieved information from records and the computers using data capture forms. Patients of service providers were not questioned or participate in the study.**
***Conclusion*:The validity of self-assessments in the HIV/ART/PMTCT Improvement Collaborative in Mtwara region of Tanzania improved as the collaborative matured. Data from computerised data bases unreliable, calling for more training in the use of computers. The weakness in communication should be addressed by collaborative designers and coaches.**

INTRODUCTION

The evidence is mixed about the impact of Continuous Quality Improvement and Healthcare Quality Improvement (QI) Collaboratives on health outcomes and the quality of care, some studies reporting positive impact, Franco *et al* 2009 (1), USAID Health Care Improvement Project, 2008 (2) and others conditional or no impact [Gordon, 1991 (3); Landon *et al* 2004(4); Mittman 2004 (5); Woolliscroft 1993 (6)]. Improvement collaboratives rely on QI teams' self-assessment of their own performance and results. Thus, the validity of self-assessment data is important to both the QI teams and the collaborative as a whole. First, self-assessment data provides QI teams with information required for them to identify quality problems, and the feedback to learn if actions they have taken are actually improving quality. If self-assessments are not valid and indicate that most patients are being given good care when in fact that is not true, then QI teams may stop seeking methods of better care in the mistaken belief they have already achieved it. Second, self-assessments that report significant improvements in healthcare quality can motivate others to adopt changes emerging from the collaborative effort, in the belief that they can do the same. Thus invalid self-assessments can spread an ineffective strategy as well as an effective one, may

contribute to maintaining ineffective strategies, and may be an important factor in determining whether an improvement collaborative is effective or ineffective.

Self-assessment is widely used in the health care field, and widely discussed in the health care literature, Bose *et al*, 2001 (7). Bandura,1977 (8) published a theory of the self-assessment process that, according to Levine, involves observation of behavior, evaluation of that behavior, and a reaction to the evaluation—more than simply measuring one's own performance, it also includes an interpretation of that performance.Marienau C (9) identified four benefits associated with self-assessment: learning from experience, functioning more effectively, strengthening commitment to competent performance, and fostering self-agency and authority. Levine E L (10) in introductory remarks for the symposium "Organisation applications of self-appraisal and self-assessment, notes that self-appraisal and self-assessment have assumed a central role in psychological research and theory. Bose *et al* (11) comment that: "All of these benefits are important in the healthcare setting. In less developed countries where resources are very limited and workers often must work on their own, the relative ease in executing self-assessment and its low cost make self-assessment especially appropriate." In a study to examine obstetrics and gynecology residents' self-assessment proficiency on a variety of surgical bench procedures Mandel *et al* (12) showed good reliability and validity of self-assessment by surgical residents; while Conroy *et al* (13) showed poor agreement between patient survey and electronic medical record. On the other hand, Hermida *et al* (13) found that both QI teams and "gold-standard" experts reported compliance highest for prenatal and immediate newborn care, and lowest for use of oxytocin, after taking agreement due to chance into account. Houston *et al* (14) conclude that residents' self-abstraction is good enough to be an alternative to costly trained abstractors.

The Government of Tanzania undertook a program that promoted health care QI collaboratives in Tanzanian health care organizations with funding from USAID and technical support from University Research Co. The program included research on the validity of QI team self-assessments in the collaboratives. Before defining specific objectives and selecting a site for the self-assessment validity study, a feasibility study undertaken by Quality Assurance Project 2008 (15) in a Tanzania Pediatric Hospital Improvement collaborative in 2006-7 concluded that: (1) a full study was feasible; (2) all written study communications with facility teams should be in Kiswahili, not English.

MATERIALS AND METHODS

This investigation was undertaken following approval by the Health Care Improvement Project and University Research Co. LLC and subsequently by the District Medical Officer In-charge of the service sites where the investigation was undertaken. All the data used for the investigation was abstracted from routine care records and therefore no human subjects were involved in the investigation. The process of abstracting the information from routine records did not influence the care processes in any way. Neither the investigators nor their agents interacted with patients in any way. For these reasons it was determined that informed consent by the health workers or patients was not required. I wish to further confirm that the data used in this study was anonymised and therefore ensured completes confidentiality and no ethical concerns.

The study investigated validity of self-assessment data produced and used by nine facility-based QI teams participating in the Mtwara region HIV/ART/PMTCT improvement collaborative in southern Tanzania (referred to hereinafter as the "ART collaborative") during its first 10 months. Quality was defined as compliance-with-standards. This study differs from other studies of the validity of QI team self-assessment in two important ways. First, it measures *change* in validity of self-assessment produced over the first 10 months of the collaborative. Second, because self-assessments are actually a composite of several self-assessment activities, it investigated the validity of each activity used by the QI teams to self-assess the quality of ART care, as shown in Box 1.

The study had four objectives:

1. Define the separate self-assessment steps that contribute to the validity of the QI Team estimates of their own performance.
2. Determine if the validity of self-assessments by the Tanzania QI Teams improved during the first year of the collaborative.
3. Determine if the pattern of validity improvement during the first year is the same for the different self-assessment activities.
4. Determine if validity of self-assessments differs by self-assessment activity, facility, or type of care.

The collaborative focused on improving the quality of care for eight different care services where the quality of care for each service was defined by a compliance indicator that measured the percentage of eligible cases who received the service according to a pre-defined standard of care. The nine QI teams in the Mtwara collaborative were trained in the improvement collaborative approach prior to the study initiation, and then received regular (usually quarterly) visits from an experienced QI clinical professional ("coach") who advised and assisted them in the QI process. The collaborative recommended

that 30 cases be randomly selected from among all patients seen during the previous month.

This study examined the validity of QI teams self-assessments for three of the eight types of care, each with its own performance indicator. The three were selected from among the eight possible types of care using the following criteria: the associated indicator was measured by all or most of the study facilities; the system for providing care included both computerized data and written record data sources; and the selected indicators included PMTCT for women, ART for adults, and services for children.

The three indicators selected for the study are:

1. Percentage of HIV-positive pregnant women attending ANC who are enrolled in the Care and Treatment Center ("PMTCT enrollees").
2. Percentage of exposed children less than 18 months who receive daily Cotrimoxazole prophylaxis ("Cotrimoxazole").
3. Percentage of HIV patients who are assessed for active TB every visit ("TB").

The study defined eight activities in the self-assessment process that can influence the validity of self-assessments by facility-based QI teams (Box 1). Failure of any of these activities can lead to inaccurate self-assessment.

Data Collection: The Mtwara collaborative began in June, 2009. The validity of the self-assessment activities was measured at regular intervals in each of the nine facilities over the first ten months of the collaborative by three of the authors referred to as "expert reviewers" hereafter. Baseline measurements in July, 2009 were obtained during a September visit to each site; and subsequent visits in November, February and June measured self-assessment validity in the month or months just preceding the visit. Measurement results were kept hidden from participating sites and their coaches by not discussing the study, the study procedures, or results until after the study was completed.

Data collection approaches differed across the eight activities, but included both objective measurements and expert reviewer subjective judgments for several activities. Some facilities used computerised databases for activities 2, 3, 4 or 5.Two different data collection forms were designed for these activities, one for facilities using a computerised process and another for manual process. In all, nine different data forms were developed, tested and used, as described below.

Quality of the patient record (Forms 1a and 1b): In Form 1a, the Expert Reviewers determined the completeness of patient records, as measured by the percentage of certain boxes and blanks that were filled in. (They did not attempt to determine whether the information written in the boxes and blanks was accurate.) In Form 1b the reviewers estimated the quality of the notes based on whether relevant information was present, when written, and who wrote them. Finally, the experts recorded their overall impression of the quality of the notes using a 5-point Likert scale.

Storage and retrieval of records- manual system (Form 2): First, the expert reviewers judged the quality of storage, including where the records were kept, whether easily retrieved, if kept in a secured room, and if adequate confidentially and how easy it was to identify and retrieve records of patients who were lost to follow up, transferred out, or dead.

On each visit the Expert Reviewer judged the quality retrieval by randomly selecting 30 cases receiving ART/PMTCT care during the previous month, and then attempted to retrieve the records within three hours. (If fewer than 30 received care the previous month, then cases in the previous two or three months were sampled, to a maximum of 30 in the review sample.) The retrieval score for the month was the percentage of the entire sample of records (30 in most months) that were received within the 3 hours.

Selecting a sample for abstracting – manual system (Form 3): To determine whether the QI team was selecting the sample correctly, the Expert Reviewers asked the QI team member responsible for selecting the sample how it was done and then requested that the QI Team member actually do the sampling procedure and observed it. A correct procedure for sampling 30 records had been previously recommended by the collaborative leaders to all QI teams. The Expert evaluated the correctness of the procedure by noting: (1) whether the recommended procedure was actually followed, (2) whether the sample was spread out evenly over the month or months, and (3) the Reviewer's overall impression of the sample selection as performed. The reviewer scored each of these three elements on a five-point scale from strongly disagrees to strongly agree.

Abstracting data from patient records – manual system (Form 4): In the Collaborative program, each QI team draws a sample of 30 records of patients who received care in the previous month, and for each patient assessed if the care reported in the record complied with the standard-for-care for the condition (YES or NO). The Expert's assessment of the same records was considered correct (Gold Standard) and compared to the QI Team's assessment of each record. In the "Gold Standard" column of Form 4, the Expert entered "YES" if the patient record indicated the care complied with the standard, "NO" if the reviewer decided the record indicated there was no compliance with the

standard, and "LOST" if the record was not abstracted or if the QI Team conclusion about compliance was not understandable to the Expert.

The performance score for the month equaled the percentage of cases in which the expert and QI team agreed. The QI team and expert were in agreement only if both wrote YES, or both wrote NO. They disagreed if the expert wrote LOST or if one wrote YES and the other NO.

Summarising abstracted data – manual system (Form 5): The validity of this activity depends on whether the QI Team summarised their own abstracts correctly, not whether the QI Team abstracts were done correctly. For example, the QI Team could abstract all the records correctly, and then add them up wrongly, or they could add up the results of their abstractions correctly even though some of the abstracts did not reflect accurately what the record said. For each indicator and month the Expert calculated the percentage of compliant cases for the month based on the abstracts done by the QI Team, and compared it to the percentage of compliant cases the QI Team recorded for the month. The Expert summarisation was assumed to be correct. If the two percentages agreed, the QI Team received a perfect summarisation score (100%). If they didn't agree, the summarisation score equaled:

100 – ((Diff / QI Team Score) x 100)

Where Diff = Absolute value of [QI Team % compliant – Expert % compliant].

Entering computer data - agreement of computer and written records (Form 6): The data entered into computerised databases was judged correct if it agreed with the original written record for Indicators 1 and 3 (This information was not collected for Indicator 2.). During each health facility visit, the Expert Reviewer determined if the computer data agreed with the written record for each of the 30 randomly selected cases from the past month (or months). The validity of the computer data for indicators 1 and 3 was the percentage of the cases for which the computer data agreed with the original record.

Computer processing – internal logic and quality (Form 7): On Form 7 the Expert Reviewers made an overall judgment about the completeness and correctness of the computer data, based on the logic and procedures used by the computer to calculate the numerators and denominators for each indicator. The judgment was quantified using a 5-point scale varying between very poor and very good.

Communicating findings (Form 8): To assess the quality of communication from individual QI team members to the other members of the team and other providers, four communication activities were investigated by the Expert Reviewer: how results were presented to the QI team; how presented to other providers; whether discussed with QI team members and other providers; and whether private conversations regarding the results were held with providers, especially those providing the relevant care. The Expert Reviewers assigned a quality score of 0-3 to each activity depending on whether: (1) no results were posted (score=0); (2) written and/or graphics were used (score=1); (3) verbal plus written and graphics were used (score=2); (4) verbal, written and annotated charts of results were used (score=3). Thus the highest possible communication score was 12, four activities each with a perfect score of 3.

Field testing data collection procedures and forms: The data collection forms and procedures were pre-tested by the Expert Reviewers in Dar es Salaam facilities not involved in the study.

Data Analysis: Data were entered and cleaned in EPI INFO 5.31 and analysed in STATA.

RESULTS

The Mtwara ART collaborative included one regional hospital, four district hospitals, one mission hospital, and three health centers as shown in Table 1.

Table 1
Characteristics of the Study Sites

Study site	Type of facility	Beds	FTEs on post	Number of women in ANC	Registered HIV+ children < 18 months	Number of HIV+ patients
1	Health center	21	13/35	238	1	323
2	District hospital	155	206/336	1,016	29	1,858 (732 on ARV)
3	Regional hospital	320	250/329	203	12	2,548
4	Health center	17	25/40	533	2	556
5	Health center	10	17/40	382	15	217
6	District hospital	165	117/229	1,227	2	919
7	Mission hospital	300	305/426	4,885	21	2,422
8	District hospital	227	176/235	2,560	6	869
9	District hospital	28	32/235	886	10	132

Table 2
QI SA Validity Improvement in 9 Facilities during First Year of Collaborative

SA Activity 1, 2 / Measurement	Indicator 3	Change in validity during study	P value
1. Record writing			
a. blanks filled A	Indicator #1	8.1% increase, nearly significant	p=0.010
	Indicator #2	Insufficient data	
	Indicator #3	11.2% increase, significant	p=0.001
b. Who wrote record A	Indicator #1	27.8% increase, significantp=0.054	
	Indicator #2	49.3% increase, significant	p=0.027
	Indicator #3	Insignificant change, high scores throughout study	p=0.089
c. When written A	Indicator #1	Insignificant small increase, periods 4-5 high (~ 0.95)	p=0.09
	Indicator #2	45.5% increase, significant	p=0.039
	Indicator #3	Insignificant small increase, periods 4-5 high (~ 0.95)	p=0.093
d. Overall procedureB	Indicator #1	Increase 2.0+ times more likely in successive period	p=0.011
	Indicator #2	Increase 3.0+ times more likely in successive period	p=0.002
	Indicator #3	Increase 1.9+ times more likely in successive period	p=0.005
2. Store and retrieve records			
a. Retrieval within 3 hours A			
	Indicator #1	Insignificant change, high scores throughout study	p=0.150
	Indicator #2	Insufficient data	
	Indicator #3	Insignificant small decrease.	p=0.663
b. Storage proceduresB	Indicator #1	No significant change in 6 of 8 storage indicators.	
		Significant increase in Flow and Arrangement indicators	
		Flow:	p<0.000
		Arrangement:	p<0.000
	Indicator #2	No significant change in 6 of 8 storage indicators.	
		Significant increase in Flow(+1.8)andArrangement(+3.3) Flow:	p<0.000
		Arrangement:	p<0.000
	Indicator #3	No significant change in 6 of 8 storage indicators.	
		Significant increase in Flow(+1.8) and Arrangement(+3.3)	
		Flow:	p=0.000
		Arrangement:	p<0.000
3. Sample selection			
a. Proper procedureB	Indicator #1	Increase 2.3+ times more likely in successive period	p=0.010
	Indicator #2	Increase 2.7+ times more likely in successive period	p=0.016
	Indicator #3	Increase 3.4+ times more likely in successive period	p=0.002
b. Was sample spread B	Indicator #1	Increase 2.1+ times more likely in successive period	p=0.005
	Indicator #2	Increase 3.2+ times more likely in successive period	p=0.018
	Indicator #3	Increase 6.1+ times more likely in successive period	P<0.000

c. Overall quality B	Indicator #1	Increase 2.3+ times more likely in successive period	p=0.005
	Indicator #2	Increase 3.2+ times more likely in successive period	p=.009
	Indicator #3	Increase 3.5+ times more likely in successive period	p=0.001
4. Abstracting quality			
a. Agree with expertA	Indicator #1	No significant change	p=0.134
	Indicator #3	44.5% increase, significant	p=0.028
5. Summarisation of abstracts			
a. Errors in summarization A	Indicator 1,2,3	No change. 8 of 9 facilities had no errors at any time	
6. Entering data in computer (agreement of computer and written records)			
Percent agreement A	Indicator #1	No significant change, small increased	p=0.168
	Indicator #2	No data - not computerized	
	Indicator #3	No significant change, small increase	p=0.123
7. Computer processing (data and logic quality)			
a. Overall judgment quality B			
	Indicator 1,2,3	Increase 2.3+ times more likely in successive period	p=0.042
b. Overall judgment of use B	Indicator 1,2,3	Increase 2.7+ times more likely in successive period	P<0.000
8. Communication of results			
a. Presentation to QI team B	Indicator 1,2,3	Increase 1.7+ times more likely in successive period	p=0.008
b. Present to all providers B			
	Indicator 1,2,3	Increase 1.7+ times more likely in successive period	p=0.003
c. Open discussion at mtg. B	Indicator 1,2,3	Increase 1.7+ times more likely in successive period	p=0.002
d. Private conversations B			
	Indicator 1,2,3	Increase 2.1+ times more likely in successive period	p=0.005

Notes:1. Analysis used STATA Linear Regression for measurements marked with superscript A, and STATA Ordered Logistic Regression with superscript B. 2. Method B results are abbreviated in the table; for example, a complete statement for results in row 1-d-Ind#1 is, "Each successive period is 2.0 or more times as likely to increase as to stay the same or decrease, than the previous period." 3. This column refers to overall performance indicators of the 3 types of care in the study (PMTCT enrollees, Cotrimoxazole prophylaxis, TB assessment).

Table 3

Self-Assessment Validity in Data Collection Periods 4+5 by Site for Activities 1-7 [1]

Self-assessment Activity	Site 1	2	3	4	5	6	7	8	9	9-Site Average
1. Record writing (ave of blanks, how, when, overall) ([AandB] Max=100)	96.3	83.1	98.1	93.5	89.1	89.4	97.4	88.6	96.6	92.4
2a. Storage ([B] Max=8) 2	75.0	100	62.5	68.8	50.0	62.5	87.5	75.0	75.0	72.5
2b. Retrieval ([A] Max=100)	100	96.5	79.0	100	99.2	89.2	91.6	75.8	100	92.4
3. Sample selection ([B] Max=5)	88.0	100	100	94.7	100	88.7	100	82.0	90.7	94.0
4. Record abstraction ([A] Max=100)	100	100	88.0	99.4	64.8	90.6	96.6	86.5	100	91.8
5. Summarization of abstracts ([A] Max=100)	100	100	100	100	100	96.8	100	100	100	99.6
6. Agreement of computer and written records 2 ([A] Max=100)	--	93.6	86.4	--	--	57.3	78.3	--	--	79.6 3
7. Quality of computer records ([B] Max=5)	--	80.0	76.0	--	--	80.0	90.0	80.0	--	81.0 3
Average of all 7 Activities (2a+2b averaged) 4	94.4	93.6	88.5	94.4	85.7	82.7	93.1	85.4	94.9	88.6

Notes for Table 3: 1.All validity scores in this table are reported as a percentage of the maximum possible score for the activity. However, validity measurements for three of the activities (2a-Record Storage, 3-Sample Selection, 7-Computer Processing) are ordinal, and therefore the percentages for these activities in the table are only approximate and are not appropriate to make statistical statements. Activities with validity measured by continuous variables are denoted by (A), and ordinal variables by (B). 2. The numbers for each site in rows 2a and 6 are the average of indicators 1 and 3 because no data was obtained for Performance Indicator #2 for these activities. 3. The "9-Site Average" for self-assessment Activity #6 includes only 4 sites and for Activity #7 only 5 sites. 4. For each site the average of 2a and 2b is calculated and entered as the value for Activity #2 in the calculation of the all-activity site average.

Communication activities: The communication activity is conceptually different from the other self-assessment activities because while the other activities, taken together, determine the information that the qi teams document as their performance scores each month, the communication activity happens after that, and may use the documentation. Also, the four validity measurements of the communication activity may not be additive. For example, it may suffice for a site to implement one or two of the communication strategies rather than all four. Nevertheless the average validity score of the communication activity was computed across all 9 sites for comparison.

Table 4

Self-assessment Scores in Periods 4+5 by Site for Communication (activity 8)

Measurement of Communication Activity	Site 1	2	3	4	5	6	7	8	9	9-Site Avarage
8a. Presented to QI Team meeting (Max=4)	0	0	50.0	75.0	0	37.5	75.0	75.0	50.0	40.3
8b. Presented in ART Providers meeting (Max=2)	0	0	50.0	62.5	0	50.0	75.0	75.0	50.0	40.3
8c. Results discussed in Provider meetings (Max=3)	0	0	33.3	66.7	33.3	33.3	66.7	66.7	100	44.3
8d. Private conversations with Providers (Max=3)	33.3	33.3	66.7	66.7	66.7	83.3	100	66.7	66.7	64.8
Average	8.3	8.3	50.0	68.8	25.0	50.0	79.2	70.8	66.7	47.5

Notes:Validity scores are reported as a percentage of the maximum possible score for the measurement.

Validity across self-assessment activities, sites or indicators: All validity scores in Table 5 are reported as a percentage of the maximum possible score for the Activity. However, validity measurements for three of the Activities (2a-Storage, 3-Sample Selection, 7-Computer Processing) are ordinal, and therefore the percentages for these activities in the table are only approximate and not appropriate to make statistical statements. Activities with validity measured by continuous variables are denoted by (A), and ordinal variables by [B]. 2. Self-assessment Activities 2a and 7 were defined as identical for all three Indicators. 3. For Indicators 1 and 3, the average of 2a and 2b is calculated and entered as the value for Activity #2 in the calculation of the 3-Indicator Average. 4. The average for self-assessment Activity #6 includes only 4 sites and for Activity #7 only 5 sites. 5. For each indicator the average of 2a and 2b is calculated and entered as the value for Activity #2 in the calculation of the indicator average.

Table 5

Self-assessment Validity in Data Collection Periods 4+5 by Indicator for Activities 1-7 [1]

Self-assessment Activity	Indicator 1	2	3	3-Indicator Average
1. Record Writing (average of blanks, who, when, overall)([AandB] Max=100)	93.6	90.4	94.5	90.4
2a. Record Storage ([B] Max=8) 2	72.9	72.9	72.9	72.9
2b. Record Retrieval ([A] Max=100)	96.0	--	88.7	92.4
3. Sample Selection ([B] Max=5)	91.9	94.4	95.2	93.8
4. Record Abstraction ([A] Max=100)	92.7	92.8	89.8	91.8
5. Summarize Abstracts ([A] Max=100)	100	98.9	100	99.6
6. Entering Data in Computer 2 ([A] Max =100)	82.5	--	76.8	79.7[4]
7. Computer Processing 2 ([B] Max=5)	81.0	81.0	81.0	81.0[4]
Average of all 7 Activities (2a+2b averaged) 5	89.4	88.4	88.3	88.7

DISCUSSION

Did the validity of QI team's self-assessments improve? The validity of self-assessment generally improved during the first year of the collaborative. Six of the eight self-assessment activities started the collaborative with low validity. Four of the six (record writing, sample selection, computer processing, communication of results) had statistically significant upward trends in the validity of the self-assessments. Two of the 6 (storage but not retrieval of records, entering data in computer), started low and did not increase. Two self-assessment activities (abstracting, summarisation) maintained high validity throughout the study. Record retrieval validity was fairly high to start with and mostly sustained good validity throughout the study. None of the self-assessment activities showed a pattern of decreases in validity during the study (tables 2 and 3.) Although the validity of communication activity self-assessments showed a small significant improvement during the study, its validity remained low at the end of the study (table 4). Changes in self-assessment validity were roughly the same for all three indicators (table 5).

How valid were self-assessment activities by the end of the study period? The average validity score in periods four and five was used to assess the validity of self-assessment at the end of the study period. Table 3 presents findings on QI team validity by the end of the study period for all self-assessment activities except communication, which is shown in Table 4. Across all indicators, sites and self-assessment activities (except Communication), overall performance in periods four and five was 88.6% of the maximum possible validity. Nine-site average validity scores above 90% were attained in the record writing, retrieval, sample selection, abstraction, and summarisation activities, while three activities (record storage, computer entry, computer processing) attained lower average validity scores. Because measures of validity included continuous and ordinal measures, the ordinal variables were transformed into continuous ones so they could be compared to the continuous variables.

Self-assessment for communication: The average communication performance score was 47.5%, substantially below the relatively high 88.6% average validity performance for the other seven self-assessment activities. Three of the nine sites (1, 2, and 5) scored low in all four validity measurements. Thus we conclude that the quality of the communication activity was relatively poor. (The ordinal measures in Table 4 have been transformed into continuous ones, with the same limitations as noted above.),

Variation across self-assessment activities: All validity scores are reported as a percentage of the maximum possible score for the Activity. However, validity measurements for three of the Activities (2a-Storage, 3-Sample Selection, 7-Computer Processing) are ordinal, and therefore the percentages for these activities in the table are only approximate and not appropriate to make statistical statements. Activities with validity measured by continuous variables are denoted by (A), and ordinal variables by (B). 2. Self-assessment Activities 2a and 7 were defined as identical for all three Indicators 3. For Indicators 1 and 3, the average of 2a and 2b is calculated and entered as the value for Activity two in the calculation of the 3-Indicator Average 4. The average for self-assessment Activity six includes only 4 sites and for Activity seven only five sites. For each indicator the average of 2a and 2b is calculated and entered as the value for Activity two in the calculation of the indicator average.

The validity scores in Table 5 show estimated validity over 90% as a percent of maximum for most activities in the last two periods of the study, but much lower (70-80%) for three activities – Record Storage (2b), Computer Record Agreement (6), Computer Processing (7). However, this lower score for three activities is less important than it might at first appear because two of the low validity activities (Record Storage, Computer Processing) are based on measurements that used Likert 5-choice ordinal variables in which the next to highest choice was 75%, and the third activity (Computer Entry, agreement of computer and written record) suffered from one especially low measured validity score.

Variation of validity of self-assessment across sites: The three lowest scoring facilities in Table 3 all had low validity scores in the Record Storage activity (2a). There were facility variations in the validity of Record Retrieval, Record Abstracting and Computer Processing. The three highest validity sites (1, 4, and 9) did not use computers for any of these activities and therefore had no validity scores for self-assessment activities 6 and 7.

Sites 1, 2 and 5 had the lowest validity in all four components of the Communication activity. Sites tended to perform most of them well or most of them poorly. This clear-cut distinction between high-validity sites and low-validity sites is not the case for any other activities with multiple components.

Variation across performance indicators: The validity scores in Table 5 did not differ between indicators in periods 4 and 5. Validity was roughly the same across the three performance indicators for all of the self-assessment activities.

In conclusions, this study defined measurements of the validity of the several activities that the Tanzania ART Quality Improvement Collaborative carried

out in order to self-assess their own performance. These activities included writing and storing patient records, sampling records to be abstracted, abstracting individual records, summarising results, and communicating results. Of the many potential improvement objectives in HIV / AIDS care that the Collaborative was addressing, three were used for this study: (1) HIV-positive pregnant women receiving ante-natal PMTCT care who are enrolled in CTC, (2) children under 18 months exposed to HIV who receivedcotrimoxazole, (3) HIV patients checked for TB every visit. The most clear-cut results were: the validity of self-assessments improved during the study, and the level of validity in the last two periods was high, with a few exceptions. This finding held across all sites and performance indicators, and was true for a majority of the self-assessment activities.Several of the activities that did not show a statistically significant increase in validity started high and stayed high.

The analysis investigated if some activities, sites, or performance indicators had low validity scores near the end of the collabrative's first year. The study found little or no difference in validity scores across sites or performance indicators,but it did identify some self-assessment activities with low validity, namely, Record Storage, and Computerised Processing.

Communication is different from the other self-assessment activities because it does not directly contribute to the information on the monthly time sequence graphs. However, communication may affect the performance improvement actions taken by providers. For example, if poor performance scores are not communicated effectively to providers, it is reasonable to think the performance will not improve. A systematic program of studies should be undertaken to learn how communication performance in collaboratives can be improved.

LIMITATIONS OF THE STUDY

1. The expert reviewers were the 'gold standard' in this study. Although attempts were made to standardize the way they arrive at conclusions, the first, second, fourth and fifth reviews were conducted by one group of experts while the third assessment was conducted by a different group of experts. Although the two groups were trained the same way and efforts to standardise them made, it is possible that inter-observer variation may have occurred. Error could also have been introduced by the frequent changes inthe QI team members during the study.
2. Although the expert visits to the sites for data collection were done as unobtrusively as possible and site staff was not informed about the study during the study, the site QI teams may nevertheless have suspected that they were being observed and may have altered their behavior as a result. If so, this could have affected the validity observed by the experts.
3. The measures of validity defined and used in this study may, in fact, not reflect true validity. Our belief is that the definitions of measurements for some activities are excellent (for example, Abstracting Records, and Summarising Abstract Results), while other definitions are without strong supportive evidence (for example, Record Writing).
4. This study measures validity over most of the first year of the collaborative; it does not address whether the trends and practices are sustained or change beyond that period.

The finding that the quality of Record Abstracting and Summarising Abstracts (activities 4 and 5) was not statistically different between QI teams and the gold standard agrees with other findings in the literature. However, our findings suggest that the usefulness of self-assessments may be compromised by other activities in the self-assessment measurement process and by the communication process.

Data provided from the computerised data bases in Mtwara may be unreliable at this stage, and less reliable than manual records. This calls for deliberate action to train and coach the QI teams in the use of computers in managing data in Mtwara.

The results of this study also show that the validity of self-assessment used in the ART Improvement Collaborative in Mtwara region of Tanzania improves as the collaborative matures. This finding coupled with the result that some activities in the self-assessment process are not done so well at the end of the collaborative suggests the need to address these activities early in the improvement collaborative.

ACKNOWLEDGEMENTS

To the American people through the United States Agency for International Development (USAID) and its Health Care Improvement Project (HCI). HCI is managed by University Research Co., LLC (URC) under the terms of Contract Numbers GHN-I-01-07-00003-00 and GHN-I-03-07-00003-00. The authors acknowledge approvals and contributions by Acting Regional Medical Officer Mtwara, Dr. Ernest Kasoyaga; the District Medical Officer (DMO) Mtwara rural and Mtwara urban Dr. R Mwakipa; and Dr. UpendoHemed respectively; DMO Masasi Dr. IgnasMlowe; DMO Tandahimba Dr. HamisMpuleni and Dr. Festo Massey; and the DMO Newala.

REFERENCES

1. Franco, L., Marquez, L., Etheir, K., *et al*. Results of collaborative improvement: Effects on health outcomes and compliance with evidence-based standards in 27 applications in 12 countries. 2009. *Collaborative Evaluation Series*. Published by the USAID Health Care Improvement Project. Bethesda, MD: University Research Co., LLC. Accessed Feb 24, 2013 at: http://www.hciproject.org/node/1397.
2. USAID Health Care Improvement Project. The improvement collaborative: An approach to rapidly improve health care and scale up quality services. 2008. Published for USAID by the USAID Health Care Improvement Project, University Research Co, LLC, Bethesda, MD. Accessed Feb 24, 2013 at: http://www.hciproject.org/node/1057.
3. Gordon, M. J. Self-assessment programs and their implications for health professions training. *Academic Medicine.* 1992; **67**: 672–679,
4. Landon, B. E., Wilson, I. B., McInnes, K., *et al*. Effects of a quality improvement collaborative on the outcome of care of patients with HIV infection: The EGHIV study. *Ann Intern Med.* 2004; **140**: 887-896.
5. Mittman, B. S. Creating the evidence base for quality improvement collaboratives. *Ann Intern Med.* 2004; **140**: 897-901.
6. Woolliscroft, J. O., TenHaken, J., Smith, J., *et al*. Medical students' clinical self-assessments: Comparisons with external measures of performance and the students' overall self-assessments of overall performance and effort. *Acad Med.* 1993; **68**: 285–294.
7. Bose, S., Oliveras, E. and Edson, W. N. How can self-assessment improve the quality of healthcare? *Operations Research Issue Paper.* 2001; **2**. Published for USAID by the Quality Assurance Project, University Research Co, LLC, Bethesda, MD and JHPIEGO Corporation, Baltimore, MD.http://pdf.usaid.gov/pdf_docs/Pnacn247.pdf
8. Bandura, A. Self-efficacy: Towards a unifying theory of behavioral change. *Psychological Review.* 1977; **84**: 191–215.
9. Marienau, C. Self-assessment at work: Outcomes of adult learners' reflections on practice. *Adult Educ Q 1999;* **49**: 135–146.
10. Levine, E. L. Introductory remarks for the symposium "Organization applications of self-appraisal and self-assessment: Another look." *Personnel Psychology* 1980; **33**: 259-262.
11. Conroy, M. C., Majchrzak, N. E., Silverman, C. B., *et al*. Measuring provider adherence to tobacco treatment guidelines: a comparison of electronic record review, patient survey, and provider survey. *Nicotine Tobacco Res.* 2005; **7**: 35-43.
12. Mandel, L. S., Goff, B. A. and Lentz, G. M. Self-assessment of resident surgical skills: Is it feasible? *Am J ObandGyn.* 2005; **193**: 1817-1822.
13. Hermida, J., Broughton, E. and Franco, L. M. Validity of self-assessment in a quality improvement collaborative in Ecuador.*Int'l J Qual Health Care.* 2011; **23**: 690-696. http://www.ncbi.nlm.nih.gov/pubmed/21840942
14. Houston, T. K., Wall, T. C., Willet, L. L., *et al*. Can residents accurately abstract their own charts? *Acad Med.* 2009; **84**: 391-395.
15. Quality Assurance Project. Feasibility of self-assessment validity study report – Tanzania June 14, 2008). University Research Co., LLC. Bethesda, MD. HCI Project Annual Report, Page 72 http://www.urc-chs.com/uploads/resourcefiles/hciyearoneannualprojectreportmasterlani21_jb22.pdf.