MEDICAL CARE
June 1984, Vol. 22, No. 6

Improving Physician Performance Through Peer Comparison Feedback

RICHARD N. WINICKOFF, MD,* KATHY L. COLTIN, MPH,† MARY M. MORGAN, AB,‡ ROBERT C. BUXBAUM, MD,† AND G. OCTO BARNETT, MD‡

A project to improve physician performance in colorectal cancer screening was evaluated as part of an ambulatory quality assurance program. A minimum standard was adopted requiring a digital examination and stool test for occult blood at annual check-ups of patients aged 40 years and older. During a 3½-year period, three different intervention strategies for improved compliance with the standard were sequentially implemented and assessed: educational meeting, retrospective feedback of group compliance rate, and retrospective feedback of individual compliance rate compared with that of peers. A pretest/posttest design was employed in evaluating the first two intervention strategies. Neither strategy resulted in significant improvement in compliance. Monthly feedback of individual performance ranked with that of peers was then implemented in a randomized clinical trial utilizing a crossover design. During the first 6-month period, the physicians receiving feedback (group 1) improved from 66.0% to 79.9% ($P < 0.001$), while the control group (group 2) also improved, from 67.5% to 76.6% ($P < 0.001$), suggesting a spillover effect. During the second 6-month period, group 2 received feedback and group 1 did not. Group 1 stabilized at approximately 80% while group 2 continued to improve from 76.6% to 84.0% ($P < 0.001$). Behavior changes persisted at 6 and 12 months after intervention. Key words: physician performance; quality assurance; feedback, peer comparison; screening. (Med Care 1984, 22:527–534)

Much attention has been paid in the past decade to methods of assessing the quality of care,[1–5] but there have been relatively few instances of demonstrated improvement brought about through quality assurance programs.[6–8] Several reports have sounded a discouraging or skeptical note.[9–13] The Harvard Community Health Plan has had an active institutional quality assurance program since October 1974 that has emphasized the improvement of clinical performance. Through application of different intervention methods, this program has succeeded in producing and documenting improvements in clinical performance in several projects.[14–17] This report concerns a project in colorectal cancer screening in which the rate of performance of a stool test for blood was improved using the technique of peer comparison feedback, i.e., feedback of individual performance ranked with that of

* From the Veterans Administration Medical Center, West Roxbury, Massachusetts.

† From the Research Department, Harvard Community Health Plan, Boston, Massachusetts.

‡ From the Laboratory of Computer Science, Massachusetts General Hospital, Boston, Massachusetts.

Supported by PHS Grants HS 02142 and HS 00240 from the National Center for Health Services Research.

Preliminary results presented at Eastern Section, American Federation for Clinical Research, October 1979.

Address correspondence to: Richard N. Winickoff, MD, Veterans Administration Medical Center, 1400 V.F.W. Parkway, West Roxbury, MA 02132.

peers. This method of intervention was developed after an initial educational meeting and subsequent feedback of group performance had each failed to bring about improvement.

Background

This project was performed in the Department of Internal Medicine at the Kenmore Center of the Harvard Community Health Plan. The Harvard Community Health Plan (HCHP) is a prepaid group practice in Boston that served 65,000 patients in two major health centers at the time of the study. The HCHP Kenmore Center provided care for approximately 35,000 of these members. The Kenmore Center utilizes a fully automated medical record information system that evolved from COSTAR.[18]

The organization of the quality assurance program and the process by which projects are implemented has been described elsewhere.[19]

Colorectal cancer screening was chosen as a quality assurance topic because of the relatively high incidence of the disease, its dire outcome untreated, and the potential of cure with early detection. The particular standard employed was that a digital examination and stool test for occult blood should be obtained at all initial or periodic check-ups of patients over the age of 40. It was recognized that this standard represented a minimum standard of high quality care, and that more aggressive screening programs had been proposed, such as periodic sigmoidoscopy[20] and multiple stool specimens obtained with the patient on a high-residue diet.[21] However, these procedures were felt to be too controversial and therefore not suitable for selection as standards.

Methods

The stool test for blood was the marker used to audit performance. It is a coded item in the automated medical record, and its presence on the date of the examination means that the physician obtained a stool specimen on rectal examination and submitted it to the laboratory. Rectal examination is not a coded item, but is entered as free text. Therefore, it cannot be easily audited.

A computer program was developed that could check all records to determine for each provider of care and for each time period the number of applicable check-ups and the rate of performance of stool tests. A time window from 6 months before to 1 month after the check-up was allowed for compliance with this standard. Compliance with the standard was 66.8% (944 of 1,413) for the entire group for a 6-month baseline period.

Three intervention methods were sequentially employed in an attempt to improve performance. These were: 1) an educational meeting at which the standard was discussed and agreed upon; 2) another meeting at which the rate of group compliance with the standard before and after the educational meeting was presented; and 3) peer comparison feedback.

The educational meeting took place during one of the regular weekly meetings of the Department of Internal Medicine and lasted about 1 hour. Members of the department were required to be at the meetings; attendance regularly exceeded 90%. At this particular meeting the rationale for screening for colon cancer was discussed, along with its epidemiology and clinical presentation. Selected papers were discussed and copies made available.[20–22] There was much interest and a lively discussion of approaches to screening. Guidelines for further evaluation of positive tests were distributed. There was universal agreement with the minimum screening standard presented above, though some internists stated that they would employ additional procedures, such as routine sigmoidoscopy. Following the educational meeting, the compliance level was 69.2% (925 of 1,336) for the contiguous 6-month posteducation period, not a significant improvement.

Because of the lack of significant improvement, a 1-year retrospective study was performed on a 10% random sample of 920 eligible records in which the stool test was not recorded. It was found that in approximately 10% of such records, there was evidence of a concomitant medical condition that rendered the test unnecessary or inappropriate. Thus, if performance and recording were brought to optimal levels, audit should show compliance levels for performing the test of about 90%.

Again, an entire department meeting was devoted to colorectal cancer. The results for the posteducation period and the results of the retrospective sample, demonstrating the 90% level as optimal, were presented. Physicians were surprised at their poor performance as a group but agreed to try harder. Unfortunately, performance for the next 2 months was 68.5% (303 of 442), essentially unchanged. In both this and the preceding analysis, there was at least a 98% chance of detecting a true difference in performance of 10% at a significance level of $P \leq 0.05$. It was felt that an improvement in compliance of 10% would be needed to be programmatically significant.

At this point, a project was undertaken to design an alternative intervention. During this project period, which lasted 14 months, an experiment was developed to test the effectiveness of peer comparison feedback for attaining the optimal compliance level.

A total of 16 internists participated in the peer comparison feedback experiment. The 16 physicians were randomized into two groups of eight, after stratification on the basis of performance during the 6-month period prior to implementing the experiment (the preintervention period). A cross-over design was employed such that the two groups received monthly feedback for alternating 6-month periods during a 12-month intervention period. The experiment was approved by all members of the department at a regular departmental meeting 3 months prior to its implementation. In addition, performance was determined for a 6-month postintervention period in which neither group received feedback. (The chi-square test was used to compare performance between groups and between periods.)

An example of the type of monthly feedback sent to physicians in the peer comparison feedback experiment is shown in Figure 1. An arrow would designate the individual to whom the report was sent. Note that individuals are identified by number, and an individual knew only his/her own number and not those of peers. Although 16 internists participated in the experiment, the computer displayed data

No. of Check-ups = 162	No. of Stool Tests Done = 114		% Done = 40.4
Provider	**No. of Check-ups**	**No. of Stool Tests Done**	**% Compliance**
1	6	6	100.0
2	7	7	100.0
3	2	2	100.0
4	3	3	100.0
5	1	1	100.0
6	2	2	100.0
7	10	9	90.0
8	17	15	88.2
9	18	15	88.3
10	8	6	75.0
11	20	14	70.0
12	6	4	66.7
13	21	13	61.9
14	5	3	60.0
15	16	9	56.3
16	4	2	50.0
17	2	1	50.0
18	3	1	33.3
19	3	1	33.3
20	7	0	0.0

FIG. 1. Individual feedback to providers.

TABLE 1. Preintervention Period

	Group 1 (No Feedback)		Group 2 (No Feedback)	
	No.	%	No.	%
Test done	549	66.0	569	67.5
Test not done	283	34.0	274	32.5
Total patients	832	100.0	843	100.0

$\chi^2 = 0.4310$ (NS).
$df = 1$.

on all physicians performing check-ups during the month. It should be pointed out that because this figure is only an example of 1 month of feedback, conclusions about performance of individual physicians or of the group cannot be drawn from it.

Results

Table 1 compares the performance for the two groups during the preintervention period. There was no statistically significant difference in the compliance levels of the two groups, but group 2 had a 1.5% higher rate of compliance.

Table 2 compares the performance for the two groups during the first 6-month period of peer comparison feedback. Both groups improved, but group 1 had a 3.2% higher rate of compliance. Again, the difference between the two groups did not achieve statistical significance. The power for detecting a 5% difference between the groups at the $P \leq 0.05$ level is 0.74, and the power for detecting a 10% difference at the $P \leq 0.05$ level is >0.99. The performance of each group considered independently did improve significantly. Group 1 increased from 66.0% to 79.9% compliance ($P < 0.001$), while group 2 increased from 67.5% to 76.7% ($P < 0.001$).

TABLE 2. Intervention Period I

	Group 1 (Feedback)		Group 2 (No Feedback)	
	No.	%	No.	%
Test done	694	79.9	750	76.7
Test not done	175	20.1	228	23.3
Total patients	869	100.0	978	100.0

$\chi^2 = 2.719$ (NS).
$df = 1$.

It is interesting that both the group receiving feedback and the group not receiving feedback improved significantly, beginning with the first intervention period of the peer comparison feedback experiment. It was hypothesized that this effect might be due in part to:

1. The effect of knowing that one's performance is being audited, whether or not feedback is given, and/or

2. The effect of informal communication with those providers receiving feedback, which should stimulate competitive drive even in the absence of direct feedback. Communication is inevitable in a group as small and close-knit as the 16-member department where this experiment was performed.

To assess these possibilities, compliance rates in both groups were compared on a monthly basis during intervention period I and for the 3-month periods immediately before and after a presentation of the monitoring program at the Internal Medicine department meeting in the middle of the preintervention period. There was considerable fluctuation in monthly compliance levels, as displayed in Figure 2. No statistically significant differences were found between the two groups either before or after the department meeting or during any month in intervention period I. However, anecdotal accounts from individual providers indicated that initial feedback reports were shared informally among department members, and the rapid improvement by group 1 in the first 3 months of intervention period I as compared with the more gradual improvement by group 2 (Fig. 2) is suggestive of an increasing spillover effect during this period. This finding has clear implications for the design of clinical trials with providers practicing in environments similar to this one.

Table 3 shows the results obtained during the second period of peer comparison feedback. With feedback, group 2 surpassed the performance of group 1 with the difference between study groups achieving statistical significance (84.1% vs. 79.8%, $P < 0.025$). Group 2 again improved its performance significantly over the prior period (84.1% vs. 76.7%, $P < 0.001$) (Tables 2, 3). However, group 1's level of compliance did not deteriorate (79.9% vs. 79.8%) during intervention period II, in spite of cessation of feedback to this group.

In the postintervention period, the performance of group 1 declined somewhat, from 79.8% to 76.3% (NS), and group 2 actually improved slightly, from 84.1% to 85.1% (NS). The postintervention period actually represents the period 7–12 months after the cessation of feedback to group 1, but only 1–6 months after its cessation in group 2.

Combined performance of both groups went from 66.7% during the preintervention period to 82.2% at its peak during the period when group 2 received feedback. This level approaches the optimal compliance level of 90% previously described. During the period in which they received feedback, all 16 internists improved their compliance with the standard as compared with their preintervention period performance. Fifteen of the 16 internists had higher levels of compliance in the postintervention period than in the preintervention period (Table 4). The one physician whose performance was slightly worse in the postintervention period was assigned to group 1 and had shown marked improvement during feedback in intervention period I, but his performance declined steadily over the next two periods without feedback.

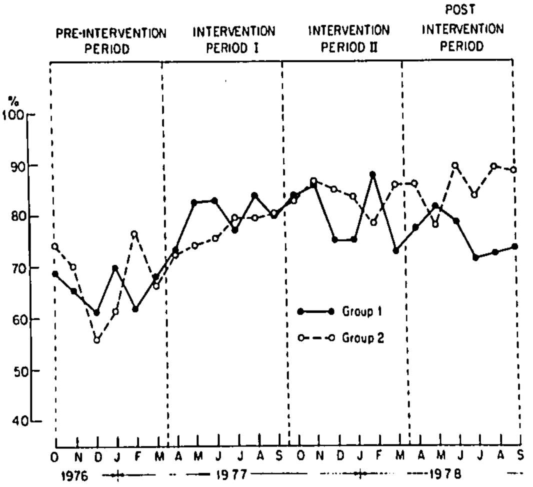

FIG. 2. Physicians' mean compliance levels (%) with care standard by study period. Group 1 received feedback in intervention period I. Group 2 received feedback in intervention period II.

Discussion

This project demonstrates the effectiveness of peer comparison feedback in improving provider compliance with a given standard of care. It should be acknowledged that the interventions that preceded peer comparison feedback were not applied in a formal experiment. Thus, the educational staff meeting and the feedback of group performance cannot strictly be compared with peer comparison feedback. However, it was because of our inability to alter performance through the continuing

TABLE 3. Intervention Period II

	Group 1 (No Feedback)		Group 2 (Feedback)	
	No.	%	No.	%
Test done	679	79.8	876	84.1
Test not done	172	20.2	165	15.9
Total patients	851	100.0	1,041	100.0

$\chi^2 = 6.0840$ (NS).
$df = 1$.
$P < 0.025$.

TABLE 4. Summary of Compliance Levels by Study Period, Group, and Individual Physician

Group/MD	Pre-intervention	Intervention I	Intervention II	Post-intervention
Group 1	66.0	79.9	79.8	76.3
MD 1	39.0	60.8	55.3	35.3
MD 2	65.5	86.8	83.4	81.9
MD 3	71.6	81.8	73.2	81.9
MD 4	50.0	81.0	71.0	53.6
MD 5	66.8	77.8	89.2	81.7
MD 6	68.9	79.9	76.1	77.5
MD 7	55.6	61.9	71.7	64.9
MD 8	81.7	88.5	91.8	87.4
Group 2	67.5	76.7	84.1	85.1
MD 9	23.1	29.8	66.7	53.8
MD 10	73.3	86.2	86.9	85.3
MD 11	73.3	84.4	88.8	87.2
MD 12	77.8	87.0	91.5	93.8
MD 13	52.3	71.2	62.5	76.2
MD 14	73.0	76.1	89.5	89.9
MD 15	79.8	90.0	84.1	87.3
MD 16	49.2	52.3	73.8	74.3

MD, physician.

education mechanism built into the practice, the educational staff meeting, that we decided to explore alternatives. It was not until the entire group was monitored individually, with one half of the group receiving individual feedback of overall monthly performance, that improvement was seen. This type of feedback forced providers to confront their individual lapses and stimulated the desire to do as well as or better than peers. Because feedback was anonymous and because the project had the prior backing of the group itself, the physicians did not resent it.

A characteristic feature of the response of individuals to data that they feel demonstrates inadequate performance is to question it. The rank-order tables were accompanied by a list of specific cases of failure to comply, with identifying information for the patients and the date of the check-up at which the test was not performed. This additional information was very helpful in establishing the credibility of the audit process.

The experiment was continued into a postintervention period so that certain conclusions could be drawn about the durability of behavior change brought about by the peer comparison feedback. Group 1 did not decline in performance in the first 6 months following cessation of their feedback. Performance did decline slightly in the second 6 months, however, but still was significantly better than before the experiment (76.3% vs. 66.0%, $P < 0.001$). How much and how fast performance would continue to decline is not certain. Group 2 actually improved slightly (85.1% vs. 84.0%) in the first 6 months following cessation of their feedback, demonstrating a persistent effect, but a second 6-month observation period was not possible for that group.

In both intervention periods, the group receiving peer comparison feedback performed better than the group without it. The difference between groups achieved statistical significance, however, only in the second intervention period. The lack of statistical significance between groups in the first intervention period may be an artifact caused by two factors:

1. The lower preintervention performance of group 1 as compared with group 2; and
2. The closeness of members within the department, leading to statistically significant improvements in both groups after the intervention was instituted compared with the prior control period.

The cost of this project was considerable. A record-by-record computerized audit of all active patients was required, taking approximately 15 hours per month of computer time, estimated to cost about $15 per hour. A number of possibilities exist for lowering the cost:

1. Improving computer capacity to reduce time requirements.
2. Using a sample rather than the entire data base.
3. Running less frequent audit and feedback cycles.
4. Combining several audit projects in each run. This can be done at modest additional cost because the bulk of time is used in accessing each record, not in performing the audit.
5. Setting "traps" in the computer's visit input sequence to create a subfile of patients 40 years of age and older who came in for a check-up, and subsequently running the audit on this subfile. This method involves a trade-off between lower costs and slower inputting of medical visit data.

This project could also be adapted to care settings utilizing a traditional paper record system by using an appointment log as the first line data source. For appointment logs where age or date of birth and reason for visit (e.g., check-up) are recorded, it would be a simple matter to create a tickler card file of applicable patients for manual chart review 1 month following the check-up visit. A peer comparison feedback report could then be compiled manually.

The peer comparison feedback approach that was successful in this experiment has also been used successfully by others[23,24] and is less dependent upon the type of record system or the specific criteria selected than upon the nature of the compliance problem. It can be used to address cost issues as well as quality issues. In this regard, Schroeder et al.[23] found that peer comparison feedback reduced outpatient costs associated with test ordering in a university medical center clinic. Both high and low utilizers of laboratory tests moved closer to the group mean, and middle cost users were unchanged. On the contrary, drug costs increased slightly following the feedback as low and middle cost users increased their costs, while high cost users were essentially unchanged. These results indicate that peer comparison feedback in the absence of agreed-upon standards has unpredictable results. Hirsch[24] also employed peer comparison feedback successfully, this time in an inpatient setting. Reductions in mean length of stay and days to improvement as well as numbers of selected tests ordered were all achieved following peer comparison feedback combined with individual chart review of appropriateness. Here, as in the present study, back-up data in the form of actual medical charts were provided to enhance the credibility of the peer comparison data.

Criteria for quality and cost-effectiveness will inevitably change with new knowledge and changing economic climate. It is useful to have methods of bringing about change that may be applicable to a variety of situations. Peer comparison feedback is one such method.

References

1. Payne BC, Lyons TF. Methods of evaluating and improving personal medical care quality: office care study. Ann Arbor: University of Michigan, 1972.

2. Sibley JC, Spitzer WO, Rudnick KV, et al. Quality of care appraisal in primary care: a quantitative method. Ann Intern Med 1975;83:46.

3. Hare RL, Barnoon S. Medical care appraisal and quality assurance in the office practice of internal medicine. Washington D.C.: U.S. Department of Commerce, 1973. (NTIS PB-237-943.)

4. Brook RH, Davies-Avery A, Greenfield S, et al. Assessing the quality of medical care using outcome measures: an overview of the method. Med Care 1977;15:(Supplement):1.

5. Williamson JW. Assessing and improving health care outcomes. Cambridge, Mass.: Ballinger, 1978.

6. Williamson JW, Alexander M, Miller GE. Continuing education and patient care research: physician response to screening test results. JAMA 1967;201:938.

7. Brook RH, Williams KN. Effect of medical care review on the use of injections. Ann Intern Med 1976;85:509.

8. McDonald CJ. Protocol-based computer reminders, the quality of care and the non-perfectability of man. N Engl J Med 1976;295:1351.

9. Nelson AR. Orphan data and the unclosed loop: a dilemma in PSRO and medical audit (Editorial). N Engl J Med 1976;295:617.

10. Eisenberg JM, Williams SV, Garner L, et al. Computer-based audit to detect and correct over-utilization of laboratory tests. Med Care 1977;15:915.

11. Kessner DM. Quality assessment and assurance: early signs of cognitive dissonance. N Engl J Med 1978;298:381.

12. Sanazaro PJ, Worth RM. Concurrent quality assurance in hospital care. N Engl J Med 1978;-298:1171.

13. McSherry CK. Quality assurance: the cost of utilization review and the educational value of medical audit in a university hospital. Surgery 1976;80:122.

14. Barnett GO, Winickoff RN, Dorsey JL, et al. Quality assurance through automated monitoring and concurrent feedback using a computer-based medical information system. Med Care 1978;16:962.

15. Morgan MM, Studney DR, Barnett GO, et al. Computerized concurrent review of prenatal care. QRB 1978;4:33.

16. Feldman J, Wilner S, Winickoff R. A study of lithium use in a health maintenance organization. QRB 1982;8:8.

17. Barnett GO, Winickoff RN, Morgan MM, et al. A computer-based monitoring system for follow-up of elevated blood pressure. Med Care 1983;21:400.

18. Barnett GO, Justice NS, Somand ME, et al. COSTAR: a computer-based medical information system for ambulatory care. Proc of IEEE 1979;-67:1226.

19. Winickoff RN, Barnett GO, Morgan MM, et al. Quality assurance in a prepaid group practice. J Amb Care M 1979;2:19.

20. Gilbertsen VA. Proctosigmoidoscopy and polypectomy in reducing the incidence of rectal cancer. Cancer 1974;34:936.

21. Greegor DH. Occult blood testing for detection of asymptomatic colon cancer. Cancer 1971;28:131.

22. Winawer SJ, Sherlock D. Detecting early colon cancer. Hosp Pract 1977;12:49.

23. Schroeder SA, Kenders K, Cooper JK, et al. Use of laboratory tests and pharmaceuticals: variation among physicians and effect of cost audit on subsequent use. JAMA 1973;225:969.

24. Hirsch EO. Utilization review as a means of continuing education. Med Care 1974;12:358.