Method of Evaluating and Improving Ambulatory Medical Care

Beverly C. Payne, Thomas F. Lyons, Evelyn Neuhaus, Marilyn Kolton, and Louis Dwarshius

The usefulness of an action-research model is demonstrated in the evaluation and improvement of ambulatory medical care in a variety of settings: solo office practice, prepaid capitation multiple-specialty group practice, and medical school hospital-based outpatient clinic practice. Improvements in the process of medical care are found to relate directly to the intensity and duration of planned interventions by the study group and are demonstrated to follow organizational changes in the participating sites—primarily managerial and support services initiated by policy decisions in each study site. Improvement in performance approaching one standard deviation results from the most intense intervention, about one-half standard deviation at the next level of intervention, and virtually no change from a simple feedback of performance measures. On the basis of these findings and other operational and research efforts to improve physician performance, it is unlikely that simple feedback of performance measures will elicit a change in behavior. However, noncoercive methods involving health care providers in problem identification, problem solving, and solution implementation are demonstrated to be effective.

BACKGROUND

Few reports are available on documented, successful interventions to improve the quality of ambulatory care. Exceptions include Helfer's

---

Research for this article was supported by the Center of Health Services Research (Grant #R01 HS 01583).

Address communications and requests for reprints to Beverly Payne, M.D., Clinical Associate Professor, University of Michigan Institute of Social Research, P.O. Box 1248, Ann Arbor, MI 48106. Thomas Lyons, Ph.D. is Associate Professor in the School of Management at the University of Michigan, Dearborn Campus; Evelyn Neuhaus, M.P.H. is a former Research Associate with The Health Services Research Center, University of Michigan; Marilyn Kolton, Ph.D. is Mental Health Program Consultant with the Alcohol, Drug Abuse, Mental Health Administration at the Department of Health and Human Services in Chicago; and Louis Dwarshius, Ph.D. is associated with the Organization for Applied Science in Society (OASIS) in Ann Arbor.

use of explicit criteria with two pediatric residents in an emergency room setting [1]; Inui's tutoring of physicians in the Health Belief Model [2] and to help control hypertension [3]; and Brook's sanctions of withholding payment to reduce parenteral administration of vitamins by New Mexico physicians [4]. John Williamson, who reported in 1967 on the difficulties of changing physician behaviors [5], has recently published impressive evidence of changes in physician performances by focusing on patient outcomes [6–9]. In these he has demonstrated a procedure of (1) implementing quality assessment projects around prioritized topics, (2) identifying health care outcome deficiencies and strengths, (3) establishing correctable causes of deficiencies, (4) effecting significant improvements of deficiencies, and (5) attributing improvements to the corrective actions taken.

In this article we present data indicating improvements in physician performance following both planned interventions and independently occurring changes in clinic management. The process of identification of a deficiency, consensus regarding its significance and correctability, action planning for correction, and remeasurement of the deficiency observed is the essence of the intervention phase of this report. In this regard, Williamson's outcome-oriented procedures are similar to our own, with our emphasis on process and outcome.

The major hypothesis of this research is that this theoretical model of action research and data feedback conceived and put into operation by Kurt Lewin and Floyd Mann [10] is powerful enough to change physicians' behavior by noncoercive methods.

METHODOLOGY

In this section we present the methods of physician performance measurement and data collection procedures; the planned interventions and the independent, organizational changes in clinic management occurring during the study; a description of the five study sites of practice and the types of physicians involved; and the final study design that emerged, along with some of its strengths and limitations.

MEASUREMENT OF PHYSICIAN PERFORMANCE AND DATA COLLECTION PROCEDURES

Criteria for diagnostic and therapeutic performance and for expected clinical outcomes were developed for ten categories of diagnoses by six

panels of physicians representing the five study sites of medical care delivery.

The categories selected for study were:

- Periodic adult medical examination
- Periodic gynecologic examination
- Periodic pediatric medical examination
- Therapeutic use of chloramphenicol, keflex, digitalis preparations, and prednisone
- Anemia
- Essential hypertension
- Chronic heart disease (arteriosclerotic, hypertensive, rheumatic)
- Vulvovaginitis
- Urinary tract infection—acute, chronic, or recurrent
- Tonsillo-pharyngitis.

The panels developed optimal criteria items necessary for diagnostic accuracy, therapeutic management, or prognostic value for a condition. The items constituting the process criteria (diagnosis and management) were weighted for importance by the physician criteria panels. The weights were distributed by areas and then by items; for example:

Vulvovaginitis

75%	Diagnosis
	History 10%
	Physical examination 15%
	Laboratory examination 50%
25%	Therapy
100%	

These criteria were reduced to category-specific data collection instruments, and an abstractor manual was prepared for the instruction of four individuals with nurse-training background. The reliability of the data collected was repeatedly tested by interabstractor agreement rates and averaged 90 percent with a range from 87 percent in gynecologic examinations to 94 percent in the use of drugs. Across pairs of abstractors, the agreement rate ranged from 88–92 percent.

Two types of measures of physician performance were used. The first was the percentage of medical records in which each criteria item was recorded for each diagnosis. These criteria items' percentages provided discrete, descriptive, specific, behavioral, and objective feedback [11] to physicians in the participating sites on their average perform-

ance in patient care. They were the primary focus of the intervention feedback seminars. However, for analytical purposes they are unwieldy. Nonparametric analyses of the data using these measures produced the same patterns of results as the second type of measure (which follows). For these reasons, analyses using these former item percentages are not reported here.

The second type of measure was a quantitative scoring of performance constructed from indexes of the percentage of criteria items. A weight was given each criteria item by the physician panels. These were importance weights reflecting the perception of the physicians regarding the relative importance of each item identified. The total weights of the items observed, divided by the total optimal weight, calculated for each patient record is the Physician Performance Index (PPI) [12]:

$$PPI = \frac{\textit{total weights of observed items}}{\textit{total optimal weights}} \times 100$$

This measure can be calculated within a diagnostic category for individual patients, or can be averaged for a site, subsite, or individual physician.

Data collection required case identification methods which varied among the sites from a logging process in one site to computerized output from a billing system.

It was planned that from each of five sites there would be approximately 150 patient records sampled from each of ten diagnostic categories. Patient listings were generated in each of the three data collection periods. A number of diagnosis-site cells had fewer than 150 cases listed or logged during the 6- to 8-week period of observation. These were selected in their entirety. When the universe exceeded 150 cases per diagnosis, a sampling rate was chosen to approximate the needed 150 cases.

Ambulatory records do not lend themselves to accurately defined diagnostic coding or to ready accessibility [13]. As in other such studies, the percentage of records abstracted was lower than those eligible for study. In this study, the percentages of records abstracted of those sampled were 65 percent, 63 percent, and 57 percent for the three data collections. Comparable reported experiences in ambulatory care studies have included 67 percent [12], 20 percent [14], 65 percent [15], and 68 percent [16].

Neither the completion rates nor the sources of noncompletion changed dramatically within the sites during the study [17]. A total of

16,153 cases were collected and identified in three separate data collection periods from the five sites of practice.

PLANNED INTERVENTIONS AND OTHER HAPPENINGS

In this section, the planned intervention attempts at the sites will be described as well as "naturally" occurring changes.

Three levels of intervention intensity were planned to assess whether physician behaviors could be changed toward greater adherence to predetermined criteria for optimal medical performance.

FEEDBACK ONLY

The first, minimal level of intervention was a straightforward reporting of the performance data results to a chief of services or chief of clinic with 1-2 hours of explanation and discussion. While this method was not expected by the research team to be very powerful in effecting change, it was used for three reasons: (1) it approximated many, but not all, Professional Services Review Organization (PSRO) quality-assurance activities [18]; (2) minimal feedback, at least, was required by the sites for their participation; and (3) it was hoped that "reporting" would act almost as a control (in the sense of very minimal intervention) and as a contrast, at least, to the other, more intensive methods.

FEEDBACK AND SEMINARS

The second level of intervention entailed reports of results not only to the chiefs of services or clinics but also to a core of physicians practicing in the settings. These physicians were brought together for two-day seminar/workshops for feedback and problem solving, away from interruptions and at least a 1½-hour drive from their practice sites. These seminars were carefully designed to make the most use of the feedback data in planning change.

Many of the problem-identification and problem-solving procedures in these seminar/workshops have been successful in other industries, also, using employee survey research data [10]. Hausser et al. provide comprehensive descriptions of this type of data feedback and organizational problem-solving activity [19].

Two sets of seminars are described. The first set occurred during July 1975, after the first data collection, and the second during January 1977, after the second data collection.

Weekend seminars were attended separately by physicians from each of three sites in July 1977; a total of 40 physicians took part in this first set. During the first day, participants were introduced to the study design and related health care research. They were then introduced to the results of one diagnosis to familiarize them with the data and methodology. After completion of a presentation on working in small, problem-solving groups, the participants were divided into such groups to review the data and identify problem areas in specific diagnoses. Each group then reported its findings to the entire seminar. In the evening, a presentation on the importance of good medical record-keeping was given by the study team.

Activities on the second day focused on using the ambulatory care results to plan for action—improving care. A presentation on "Effective Action Planning" prepared the group to develop formal plans. Small groups, each formed on the basis of physician specialty, reviewed respective diagnoses for their specialties and developed action plans for improving care. Action plans thus developed included activities such as: developing history and physical forms, improving laboratory procedures, educating other staff members, monitoring the treatment of certain cases, and using allied professional staff to collect information. The specialty groups shared their plans with the other physicians from their respective practice sites, emphasizing responsible follow-through on action plan implementation.

The second set of seminars was held during January 1977. All physicians who had been participants in the first set were invited to return, and other physicians, who had not been involved previously, were asked to attend. Thirty-nine physicians attended these seminars, with low attendance in the last two weekends due to Midwest blizzards. A special additional series of workshops was held in March and April for those 21 physicians who had been prevented by snow from attending the regular sessions.

The purpose of the second seminar was similar to the first: (1) to share results—this time, of the second data collection, and (2) to plan for improving care based on physician performance. The second series of seminars differed from the July 1975 sessions in that data were available on changes that had taken place after that first conference. The changes in physician performance could be related to action plans that had been developed by participants in the first year's sessions. If change was not present in a diagnosis, or if physician performance actually had declined, reasons for these developments could be considered in terms of action plan follow-through for these diagnoses.

The first day of the latter seminar began with an introduction to

its purpose: data review and planning for change. Key representatives from practice settings were asked to discuss changes which had been made in their respective sites as a result of the first seminar. In reviewing these changes, the participants discussed the previous year's action plans and steps they had or had not taken to implement these plans. Based on their reports, organizational factors which served to help or hinder the change process were summarized. The review of these organizational dynamics helped pinpoint reasons why changes had or had not occurred at the site. An example diagnosis was presented to the participants to show how the change data from 1975 to 1976 were recorded and how the data would be reviewed in the small groups. Members of the study team then divided participants into small, specialty-area groups, and acted as group leaders. In each case, there were generally two groups of internal medicine specialists, one group of pediatricians, and an obstetrics/gynecology group. The small groups were asked to review certain diagnoses related to their specialty: (1) the item percentages for each diagnosis, (2) identification of those items in which significant positive or negative change had occurred, and (3) identification of those items in which change was still needed. The group continued on this task for the rest of the morning and most of the afternoon. In late afternoon, additional results were presented from the study. The evening session introduced the participants to the results of a pilot study of physician and consumer attitudes toward medical care.

The second day of the conference focused on action planning. The session began with a presentation on considerations for planning for change. This presentation reviewed the factors that participants must consider in developing effective action plans. The physicians then worked within their specialty groups again, to develop action plans for the diagnosis they had reviewed on the previous day. Most of the morning was spent in action planning. In the afternoon, results of the action planning were presented to the total seminar as a means of sharing changes that each group intended to make.

ONGOING FOLLOW-UP CONSULTATIONS

The third level of intervention involved the same as the previous two but added continuing postseminar consultations on implementing changes for improved performance. These later consultations consisted largely of (1) meetings of research team members with medical directors and hospital administrators to assist them in planning methods to maximize staff involvement in creating change; (2) help by the research team in designing special sessions with primary care depart-

ments to summarize the major deficiencies identified at the seminars as well as the changes needed and action plans for their implementation; and (3) attendance by members of the research team at regularly scheduled department meetings, during which progress with implementation plans—and the problems confronted—were reviewed and discussed. Between the first and second seminar, these interventions focused largely in the pediatric department. After the second, in which the greatest improvements were found in pediatric diagnoses, the frequency and extent of these consultative interventions increased for the other departments as well.

The choice of control and intervention sites was determined before the first data collection took place. The choices of sites were not random; they were chosen for study at a particular intervention level partly on the basis of geography and cost, partly on their expressed willingness to participate in the interventions. Two sites were chosen to serve as the control—one university site and half of a solo practitioner site—while the second university site and the other half of the solo practitioner site received the moderate-level seminar intervention. Similarly, a fee-for-service group practice site received only data feedback, while the prepaid group practice site received the more intensive combination of data feedback, workshops, and follow-up consultations.

ORGANIZATIONAL CHANGES IN CLINIC MANAGEMENT

One of the problems with field experimental or quasi-experimental designs such as this is that so-called "control" sites do not maintain their status quo on all relevant variables, especially over a 3-year period. One of the two control sites underwent relatively little change, while the other one experienced management changes in nearly all of its clinics. In the first control site a younger physician, who had also participated in the criteria-setting for the study, replaced an older chief of service. In the second control site, however, the following changes occurred: a specialty hypertension clinic, which had just started when the baseline measures were collected, built up an increasing number of referrals from its hospital's main clinic; a new full-time position of medical director of one of the clinics was created and filled, with the director's office physically located in the clinic. Additional support was afforded by the creation of a full-time position of clinic business administrator to serve as assistant to the new medical director; in another of the clinics, a committee chairman was employed full-time in the clinic to act as chief, succeeding the previous nominal director whose time had been split between his academic chairmanship and his clinic office.

Only one of the clinics in this site did not undergo this type of change in clinic supervision. Over a 3-year period, even the intervention sites experienced some changes in management of some of their clinics.

The site with the most intensive intervention efforts had few changes that were not related to the action plans developed at the workshops. One of these was the replacement, after the second data collection, of a chief of service in a department with some degree of internal turmoil. In the "seminar intervention" university hospital, one of the outpatient clinics had been managed by ten subspecialty faculty members, each of whom had put in a half-day per week supervising care delivery. After the first data collection, this management arrangement was changed by hiring a full-time director who set up his office in the clinic. He was hired because of his reputation of commitment to high-quality ambulatory and preventive medical care.

In each instance, these changes were unplanned by the research team and unrelated to the study in that they were not instigated in reaction to the study results or interventions. Also, and perhaps quite importantly, each of these changes represented the allocation of increased organizational resources to the clinics in terms of medical supervisory/management time and effort brought to bear on outpatient care delivery. This was done either formally—as in the hypertension clinic, or in the changing of the chiefs' positions from part-time to full-time and increasing administrative staff support—or informally, in terms simply of more hours for new chiefs in ambulatory care settings.

SITES OF PRACTICE

Five sites of practice were studied. The first, referred to as the *Solo Practice site*, was composed of primary care physicians practicing in one Midwest county. Their practice settings included solo practices, partnerships and small groups, and a small multispecialty group. During the three data collection periods, a total of 71 physicians participated: 29 internists, 19 general and family practitioners, 11 pediatricians, and 12 gynecologists. A core group of 13 internists, 5 family practitioners, 5 pediatricians, and 7 gynecologists continued participation throughout the 3 years. Of these 30 who were able to participate in all three data collections, 16 were involved in the seminars.

The second site, referred to as the *Experimental University site*, was a teaching hospital, providing both tertiary care to the region and primary care to the local area. The primary distinguishing characteristic of this hospital and its outpatient department clinics was their involvement in teaching and training medical students, residents, nurses,

physical therapists, and other health professionals. In this site, cases abstracted for study had been attended by a total of 392 different physicians over the three data collections: 202 internists, 1 family practitioner, 45 gynecologists, 67 pediatricians, and 77 other specialists. There were 118 physicians in the last data collection, and 53 of these had had cases in data collections before the seminars had convened. From this site, 21 physicians participated in the seminars.

The third site, *Control University site*, was actually composed of several patient-care units, all linked in some way with a university school of medicine. One of these units was composed of physicians set up originally as a model family practice unit to train residents in the university's department of family medicine. Another unit, a large hospital that acted as a teaching hospital, provided cases from its medicine and gynecology clinics. Pediatric care was reviewed at the outpatient clinics of a separate large hospital devoted entirely to pediatric care. In all, a total of 243 physicians provided cases over the three data collections: 56 internists, 29 general and family practitioners, 55 gynecologists, 88 pediatricians, and 15 other specialists. There were 102 physicians in the last data collection, and 21 of these had had cases in previous data collections. No physicians from this site were involved in the seminars.

The fourth site, *Control Group-Practice site*, was a large, metropolitan, fee-for-service, teaching hospital staffed by a group of physicians representing all specialties, with all physicians on salary. They were providing extensive outpatient services, functioning as a primary care center for many patients and as a referral center for a wider geographic area. A total of 297 physicians provided cases over all data collections: 169 internists, 21 gynecologists, 38 pediatricians, and 69 other specialists. Of the 171 physicians in the last data collection, 87 had provided cases in previous data collections. None of the physicians were involved in the study seminars.

The fifth site, referred to here as the *Intensive Intervention site*, was also a group practice, operating on a prepaid basis. It was composed of a medium-sized hospital and outpatient department in the inner city, another small hospital and outpatient department in a suburban location, and three small health centers. A total of 132 physicians from this site provided data: 68 internists, 1 family practitioner, 15 gynecologists, 13 pediatricians, and 35 other specialists. Of the 83 physicians in the last data collection, 49 had had cases in previous periods. A total of 21 physicians from this site were involved in the study seminars. This was the site in which additional follow-up work was conducted by the research team.

All physicians in the Solo Practice site volunteered themselves for participation in the study and some also volunteered for the seminars. In contrast, patient records were "volunteered" in the institutions for retrospective review by chief medical officers or executive committees of the medical staff, in many cases without awareness by the attending physician. In the experimental institutions, too, some element of the seminar participants having been "volunteered" was evident.

STUDY DESIGN AND THREATS TO VALIDITY

The basic design of the study was to obtain baseline measures of performance in all sites, to intervene twice with seminars and/or follow-up in three experimental sites, and to collect performance data following each of the intervention periods in all sites. The baseline observations were made in October 1974–January 1975. The first seminars were held in July 1975. Second data collections were made in September–December 1976, followed by second seminars in January 1977 (and subsequent make-up sessions in March and April for second-seminar participants delayed by winter storms). Final data collections were made in June–August 1977. Follow-up consultations, which began in the Intensive Intervention site after the first seminar, continued up to the last data collection.

Campbell and Stanley [20] have the best presentation of the various threats to both internal and external validity in experimental and quasi-experimental research designs. Using their widely accepted terminology, the study from which this article originated is an example of what they call a "patched-up design" (p. 57). The study is described most accurately as a mixed bag of several longitudinal, quantitative case studies. It is, quite frankly, untidy and inelegant, and the results are not very convincing from any particular site or subsite. However, results from the variety of subsites, each with different design strengths and weaknesses, when combined and considered together, do suggest strongly that both increases in clinic management and the more active interventions did precede improved performances to a greater degree than where they were not present. A few brief comments about some of the Campbell and Stanley threats to validity follow.

It will be obvious from the tabular results that a general rise in performances over time was not an example of instrumentation decay: in each of the diagnostic groupings the average performance did not change, and in some sites or subsites it actually declined. Data acquisition did improve in some sites or subsites in the second and third data collections, but these improvements in data acquisition were not followed systematically by higher PPIs.

In many institutional subsites, even in those with fairly dramatic improvements in performance, most of the physicians were unaware of the study team. No opportunities were presented in these sites for either the effects of testing or of test reactivity to operate where the data abstraction of the recorded physician behaviors was considered to comprise the testing. In the Solo Practitioner site, data collection procedures were much more obtrusive; but even in that situation, no overall changes were observed among participants who had not attended the seminars.

A general pattern of more improvement and initially lower performance was evident among those sites or subsites which had become either management-change subsites or high-intervention subsites. However, inspection of the results suggests that the possibility of statistical regression effects is not a serious threat. Many of the highest-performance subsites maintained or increased their levels of diagnostic-group performances—e.g., the Intensive Intervention site (pediatric examinations) or the Control Group-Practice site (adult examinations)—and some sites with the initially lowest performances in a diagnostic grouping showed little improvement (e.g., urinary tract infections in the Control University site).

Of the threats to validity listed by Campbell and Stanley, the most relevant one, least countered by some elements of the "patched-up" combination of designs, is the possible influence of the interaction of site or physician self-selection into intervention-level groups. In the solo practitioner setting, physician self-selection entered into both the study and the seminars. The possibility exists that only those physicians who respond to an invitation to participate in a study such as this, and who, in addition, respond to a seminar invitation, will improve their performances following such seminars. A related proposal to investigate motivational, attitudinal, belief, and value differences among these physicians to study this and other questions was approved but not funded. A true random-assignment design is necessary to answer this question definitively, particularly for solo practitioners.

Self-selection into intervention groupings on the individual-physician level is evident in the Solo Practice site. Self-selection by individual physicians in the institutional settings is not as clear-cut. The decision that a site would participate in the study and/or at the seminar/follow-up levels of intervention was made by the chief medical officer of the site or by the executive committee of the medical staff. In the institutions, the performances of the individual physicians were "volunteered" by their medical officers, in contrast to the Solo Practice site, where each physician decided for himself whether to participate in

the study or not. In the experimental institutions, even seminar participation was somewhat "volunteered" to varying degrees, again in contrast to solo setting participation.

The "volunteered" nature of the performances of individual physicians in the institutional settings, coupled with the lack of improvement of nonseminar participants in the solo setting, suggests that individual self-selection into the study itself may not pose a serious threat to our conclusions. The effects of self-selection into the seminars, however, does remain a possible threat to our conclusions; that is, it may be possible that postseminar(s) improvements in performance result from the interaction of the seminars with some special characteristics of the nonrandomly assigned participants who volunteered or who did not successfully avoid being volunteered to attend and plan improvements.

However, the "volunteered" physician who attended the seminars could not alone have accounted for the performance change in the second and third data collection.

In the Experimental University site, there were 21 seminar attendees of the 392 physicians who contributed to the observed care (Table 1). In the Intensive Intervention site, 21 of the 132 physicians who contributed to the observed care attended the seminars. The task accepted by these seminar participants was to disseminate the performance record of the site to their colleagues and to induce compliance with the action changes developed to improve performance. The positive changes observed are thus explained by the "yeast" effect and the supervisory role of the seminar participants in the individual sites.

An additional design limitation of the study concerns the external validity or ability to generalize the findings. The majority of the data were collected from clinics of institutions, most of which were teaching and/or university hospitals. Also, two of the institutions were group-practice settings. The design, and these sites, were chosen to maximize the range and diversity of settings rather than to provide a representative sample of any one type of care-delivery site. As such, generalization of results beyond these particular sites is unwarranted, particularly with regard to site comparisons.

DATA ANALYSES

A Physician Performance Index (PPI) was the major measure used for each participating physician in each diagnostic category. Changes in PPIs over the three data collection periods in each site were calculated and tested for statistical significance.

The average levels of PPIs differ from diagnosis to diagnosis, and

Table 1: Design and Characteristics of the Sites

Site	*Nature of Participation and Data Collection Obtrusiveness*	*Intervention Level*	*Numbers of Physicians:* *In Study*	*In Last and Previous Data Collections (Before/After)*	*In Seminars*	*Changes in Clinic Organization*
Solo Practice	SF,* Hi 0‡	Feedback only	55	14	0	—
		Feedback and seminars	16	16	16	—
Experimental University	V'd,† Low 0‖	Feedback and seminars	392	53	21	OB-Gyn Clinic gets full-time director with office in clinic area
Control University	V'd, Low 0	Feedback only	243	21	0	Hypertension Clinic operating; Pediatric OPD gets full-time director with office in clinic and full-time business administrator; Medical Clinic gets full-time director
Control Group Practice	V'd, Low 0	Feedback only	297	87	0	Pediatric Clinic Chief changes
Intensive Intervention	V'd, Low 0	Feedback, seminars, and follow-up	132	49	21	OB-Gyn Department Chief changes

*SF = Self-volunteered.
†V'd = Volunteered by institution.
‡Hi 0 = Highly obtrusive data collection.
‖Low 0 = Lowly obtrusive data collection.

the variances of the PPIs differ among diagnoses. For analyses where PPIs are combined across diagnoses, the PPIs are standardized (*Z*-scored) within diagnostic subcategories before they are combined. The resulting standardized scores indicate the relative position of each performance in a distribution of values within a diagnostic category; each distribution has a mean of 500 and a standard deviation of 100.

For all analyses in which PPIs were used, nonparametric analyses were also made using individual criteria item percentages. Their results duplicated the results using the PPIs and are not reported here.

RESULTS

In this section we present data concerning changes in the performances of physicians in the study sites over the three data collection periods.

Average PPI levels of each site and for each data collection are presented in Table 2 for each of the ten diagnostic groupings for 1974/75 and 1976, and for each of the nine groupings in 1977. Drug management data were not collected in 1977 due to budget reductions. Statistically significant changes in average PPIs between congruent years were indicated by asterisks ($p < .05$), with daggers appearing between the yearly averages. A statistically significant difference between a 1974/75 average PPI and a 1977 average PPI is indicated by an asterisk or dagger *after* the 1977 average.

From 1974/75–1976 the Intensive Intervention site shows the greatest increase in average PPIs, increasing significantly in five of the ten groupings with nonsignificant increases in the other five. The Control University site is next in increases with four significant and five nonsignificant increases. The Experimental University site has four significant and another four nonsignificant increases. The Control Group-Practice site has three increases and one decrease that are statistically significant and another four decreases and two increases that are nonsignificant. The Solo Practice site has a balanced pattern of five increases and five decreases, with one of each showing significance.

From 1974/75–1977, the Intensive Intervention site again displays the greatest increase in performance averages. All nine of the average diagnostic PPIs are higher in 1977 than in 1974/75, and all of these increases are statistically significant. The next greatest increase is in the Control University site with seven significant increases, one nonsignificant increase and one decrease. The Experimental University site has four significant increases in PPIs, one significant decrease, and another four nonsignificant increases. The Control Group-Practice site increased average PPIs in three categories significantly, increased

Table 2: Physician Performance Indexes and Numbers of Cases (*n*) by Diagnostic Grouping, Site, and Year

Dx Group	*Solo Practice Site*			*Experimental University*			*Control University*		
	1974/75	*1976*	*1977*	*1974/75*	*1976*	*1977*	*1974/75*	*1976*	*1977*
Adult periodic examination	58 (148)	−†51 (141)	†56 (160)	70 (71)	72 (105)	−†61† (40)	18 (21)	+ †51 (5)	45† (16)
Gynecology periodic examination	77 (134)	78 (156)	77 (145)	75 (135)	77 (141)	†83† (152)	74 (154)	71 (32)	74 (26)
Pediatric periodic examination	76 (132)	78 (122)	75 (137)	79 (69)	+ †88 (65)	89† (151)	78 (148)	+ *83 (172)	83* (127)
Drugs	84 (164)	87 (86)	—a.	64 (65)	+ †82 (103)	—	82 (2)	+ *90 (38)	—
Anemias	56 (23)	56 (20)	74 (17)	60 (31)	59 (43)	72† (42)	41 (162)	48† (116)	53† (105)
Hypertension	65 (118)	68 (150)	70 (143)	73 (104)	68 (143)	*74 (206)	67 (55)	71 (133)	77* (122)
Heart disease	64 (127)	61 (140)	†67 (124)	74 (86)	75 (147)	78 (153)	58 (26)	63 (66)	*71† (64)
Vulvovaginitis	40 (61)	+ †55 (90)	*64† (86)	51 (66)	+ †72 (108)	76† (74)	29 (25)	41 (47)	†66† (60)
Urinary tract infection	66 (77)	62 (60)	65 (70)	64 (43)	67 (95)	69 (71)	49 (7)	53 (30)	53 (13)
Tonsillitis and pharyngitis	57 (117)	56 (215)	58 (150)	65 (117)	+ *72 (216)	69 (165)	52 (163)	55 (214)	†68† (89)

Dx Group	Control Group-Practice			Intensive Intervention		
	1974/75	*1976*	*1977*	*1974/75*	*1976*	*1977*
Adult periodic examination	76 (111)	+ *82 (182)	80 (158)	43 (143)	+ †63 (141)	62† (117)
Gynecology periodic examination	77 (135)	-†72 (153)	†77 (150)	70 (103)	72 (137)	†83† (165)
Pediatric periodic examination	76 (149)	+ †83 (145)	87† (204)	83 (91)	+ †89 (147)	91† (190)
Drugs	84 (35)	83 (59)		77 (104)	+ *84 (81)	
Anemias	68 (136)	72 (134)	70 (92)	44 (80)	+ *52 (171)	57† (92)
Hypertension	69 (139)	74 (132)	76 (133)	61 (147)	64 (150)	†76† (125)
Heart disease	75 (154)	74 (160)	76 (110)	55 (93)	58 (133)	†69† (152)
Vulvovaginitis	70 (115)	62 (62)	67 (72)	41 (6)	53 (164)	†67* (201)
Urinary tract infection	62 (103)	59 (71)	-*55 (79)	49 (94)	54 (122)	†62† (100)
Tonsillitis and pharyngitis	51 (59)	57 (87)	61* (119)	35 (132)	+ †46 (186)	†61† (134)

a. = Use of drugs omitted in 1971.
* $p < .05$. (See text for comparisons tested.)
† $p < .01$.

somewhat in another three, decreased nonsignificantly in two, and showed no change in one. The Solo Practice site again shows the least movement, with two significant increases, four nonsignificant increases, and three nonsignificant decreases.

Table 3 summarizes the result for standardized physician performance levels when the site data are classified by both intervention levels and changes in setting management.

The average amounts of change from 1974/75–1977 in standardized physician performance scores (Z), which correspond to the management-change/intervention totals in Table 3, are tabulated and presented in Figure 1. The same pattern of comparisons is evident: More intervention activities are followed by increased physician performances for settings both with and without management changes; and management changes precede improved physician performances overall and within each grouping of settings by intervention level.

Site A: Solo Practitioners. The Solo Practice site is the one site that has no organizational or institutional entity. It is a geographically defined set of individual practitioners who are unconnected with one another; who have no common outpatient facilities or organizational structures; who have no routine or systematic sharing of patients, medical records, nursing staffs, or ancillary personnel; whose quality of ambulatory professional performances are not systematically surveyed; who do not identify themselves and are not identified by others as a group of interacting and/or interdependent professionals. In short, they are a collection rather than a group of practitioners and, in this important regard, differ from practitioners in the other sites of practice.

A primary implication of this for the study team was the fact that opportunities are reduced substantially for widespread, systemic change interventions of technology or structure [21]. For example, instead of providing one incubator for cultures in one clinic area for use by numerous physicians, several separate incubators must be provided, each for use by only one solo practitioner.

Another implication is that while each physician has more autonomy in his/her own practice than an institutional counterpart has, he or she also is the only change implementer, and neither problems nor solutions can be divided for a reasonable, shared workload of change-action steps.

A final implication is that no ongoing, routine opportunities exist for the solo practitioners to provide support for continuing action steps, for reinforcing change attempts, or for reinforcing changes accom-

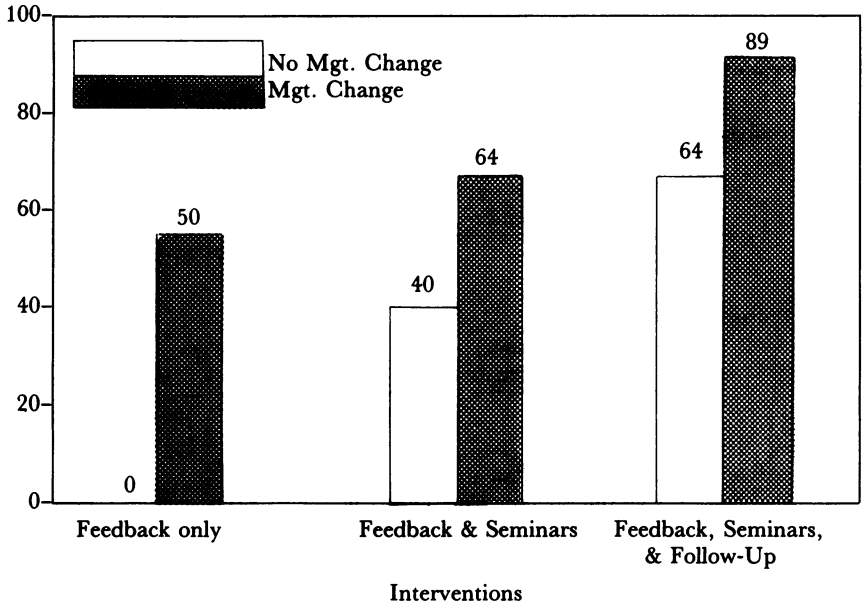
Table 3: Standardized Physician Performance Index Scores (Z) and Numbers of Cases (n) by Site, Clinic Management Change, Intervention Level, and Years

	1974/75	*1976*	*1977*
No Clinic Management Change:			
I. Feedback only:			
Solo Practice (No Seminar)	499(575)	481(691)	466(514)
Control University	447(293)	465(197)	494(113)
Control Group-Practice	528(970)	528(1,024)	533(822)
Average	506(1,838)	504(1,912)	506(1,449)
II. Feedback and seminars:			
Solo Practice (Seminar)	472(670)	494(553)	521(518)
Experimental University	505(602)	526(1,016)	535(833)
Average	489(1,232)	514(1,560)	529(1,361)
III. Feedback, seminars, and follow-up:			
Intensive Intervention	451(909)	481(1,237)	515(949)
Clinic Management Change:			
I. Feedback only:			
Control University	462(536)	487(678)	506(509)
Control Group-Practice	454(197)	502(196)	516(295)
Average	460(733)	490(874)	510(804)
II. Feedback and seminars:			
Experimental University	484(733)	521(219)	548(221)
III. Feedback, seminars, and follow-up:			
Intensive Intervention	448(112)	471(213)	537(327)

plished, as there might be in the institutional settings with regular department meetings and both formal and informal interactions.

In the institutions, the study focuses were the patterns of performance by groups of physicians reacting to systemic and commonly shared structural, technological, and, perhaps, educational change interventions in their settings. Of purely logistic necessity—and especially in the larger institutions, many of these physicians were not expected to have any direct contact with the study team. In the 71 physically and geographically separated offices of the solo site, few

Figure 1: Standardized Score Changes, 1974–1977, by Interventions and Management Change



opportunities were available for such systemic or common interventions, and none were expected to affect the performances of those physicians who were not in direct contact with the study team in the seminar/workshops.

The Solo Practice site provided the opportunity and the analytical necessity to examine separately the performance of those physicians who participated in the seminar/workshops and action planning for their own performance improvement and the performances of those physicians who were able to, or who chose to receive only the report feedback with no direct involvement in the seminar/workshops.

One of the difficulties commonly encountered with longitudinal studies of volunteer participants in panel or cohort group designs is the loss of participants ("dropouts") due to a variety of reasons such as death, illness, moving, or unwillingness to continue. Another possible problem is the addition of new members ("latecomers") to the study panel or cohort who may differ from the original panel members.

In this type of situation, one can include all performances of all physicians in each of the time periods and accept the reality that different personnel do enter and leave any setting over a period of time. One also accepts that change in average site performance may be the result

of this influx and outflow of different personnel, or of changes in the performances of the stable personnel, or of both. Another alternative is to restrict the analysis to only those personnel who are panel survivors, that is, those who are in the study from beginning to end. This alternative reduces the ambiguity about changes in mean performances. But where there is a good degree of fluidity in panel membership, it reduces also the number of participating physicians and performances (cases) and restricts the description of performance changes to that subset of practitioners who remained in the study throughout.

Both methods of analysis are presented and, fortunately, the same conclusions can be drawn from each analysis regarding the effects of the seminar/workshop interventions.

Among the 71 primary-care solo practitioners involved in the study over the four years were 30 who were able and willing to participate for all three data collections. Sixteen of these provided 1,243 cases for longitudinal analyses and were involved in one or both of the seminar/workshops. Fourteen provided an additional 776 cases over the three data collections but were not able or chose not to attend the seminars. These 30 physicians provided, respectively, 60 percent, 53 percent and 85 percent of the patient cases analyzed in the three data collections.

Fortunately, the data collection procedures allowed the matching of patient case performances with unique but anonymous physician codes. This permitted us to observe a significant increase in performance for seminar participants, with no change for the nonseminar panel survivors over the three data collections (described in more detail in Table 4).

Most of the physicians in the nonsurvivor, changing-composition group were not seminar participants. This, and the combination of high performances by early dropouts and lower performances by latecomers, helps explain the otherwise curious finding of an apparent decrement in the performances of nonseminar participants when all physicians' cases are analyzed (Table 3).

PANEL SURVIVOR ANALYSES

Of the ten diagnostic groupings presented in Table 4, the performance of the physicians not participating in the seminars decreases significantly in one condition (pediatric examination) and increases significantly in another (vulvovaginitis). The average relative performance (Z_A, PPIs standardized within this site) across all diagnoses and conditions shows a nonsignificant dip during the second data collection and

Table 4: Solo Practitioners, *Panel Survivors*: Comparisons Between Seminar and Non-Seminar Participant Performances Over Data Collections

	Non-Seminar Participant Performances				*Seminar Participant Performances*			
	1974/75	*1976*	*1977*	*1977–1974/75 (Significance)*	*1974/75*	*1976*	*1977*	*1977–1974/75 (Significance)*
Adult periodic examination	50 (26)	43–* (23)	51 (67)		56 (67)	53 (46)	59 (69)	
Periodic gynecology examination	78–* (59)	71 (58)	75 (55)		74 + † (42)	86 (56)	85 (60)	+ †
Pediatric periodic examination	77 (13)	70 (10)	59 (46)	–†	63 + † (48)	78 + † (50)	87 (83)	+ †
Anemia	65 (6)	92 (1)	40 (1)		45 (8)	56 + * (6)	75 (14)	+ †
Hypertension	63 (23)	72 (40)	71 (42)		62 (40)	64 (40)	68 (72)	
Heart disease	55 (21)	58 (12)	65 (34)		60 (33)	68 (30)	70 (68)	+ *
Vulvovaginitis	39 (21)	45 + † (31)	68 (53)	+ †	39 + * (24)	57 (33)	58 (28)	+ *
Urinary tract infection	65 (19)	59 (12)	58 (29)		64 (44)	60 + * (23)	73 (34)	+ *
Tonsillitis-pharyngitis (adult)	48 (11)	44 (8)	47 (14)		50 + * (19)	65 (46)	70 (44)	+ †
T&P (pediatric)	46 (10)	47 (8)	44 (23)		52 (23)	61 (42)	61 (46)	
Total Z_A	483 (209)	471 (203)	485 (364)		476 + † (353)	513 + † (372)	533 (518)	+ †

*$p < .05$ (See text for comparisons tested.)
†$p < .01$.

a return to almost exactly the same level in the third data collection.

In contrast, the seminar participants show significant gains from the first to the second data collection overall and in four of the ten separate categories. Additional gains from the first to the third data collection, both overall and in seven of the ten categories, also are recorded.

Additional analyses were performed using nonparametric methods [22], with the unit of analysis either: (1) the physician's average performances in separate diagnoses for which patients were seen both in the first and the last data collection periods or (2) the average relative performance across all diagnostic categories (standardized scores). Both types of analyses demonstrated the above results—that seminar participants improved their performances significantly more than did nonparticipants.

DISCUSSION

We contend that significant improvements in physician performance followed both changes in clinic management and intervention through seminar and seminar/follow-up. Possible alternative interpretations were suggested in the design section, but many of these may be discounted by the results of this untidy but wideranging, quasi-experimental design. However, true experimental randomization of sites and physicians to design conditions should be the next research step, if that is possible.

Few of the physicians studied in the institutional setting institutions were involved in the seminars (Table 1). If, indeed, the seminars were successful in improving the performances even of physicians who did not attend them, it is likely that these improvements were the result of systemic changes in the institutions and clinics. A stated assumption of the seminar staff was that performance deficiencies in these common diagnoses probably were not due to lack of physician knowledge or expertise. Rather, it was assumed that improvements might come from identifying and removing barriers in the practice settings that prevented good care performance or made it difficult. The seminars focused on identifying such problems, and the staff encouraged action planning to change and modify the clinic or office environments—the equipment needed, forms and procedures, task assignments, communications with other departments and laboratories, and patient follow-up procedures. Many, although not all, of the action steps were of this type. As a simple example, providing microscopes in a clinic where there had been none preceded a remarkable increase in microscopic

examinations of smears for the diagnosis and treatment of vulvovaginitis.

It is tempting to add our findings about changes in clinic management to the impressive sociological literature on management succession. Management succession, and its effects on organizational effectiveness and performance, has a long and respected history of investigation. The improvements in performances documented here support some of the management succession research, e.g., Torrance and Guest [23–30]. However, they do not agree with the majority of the findings, which indicate decreases in performance or effectiveness following management succession [31–33].

An alternative interpretation is suggested by a closer examination of the clinic management changes cited here and by some familiarity with the clinics. The performance improvements may reflect the fact that these clinic management changes were changes not only in supervisory personnel but in the very structure and amount of activity of the positions. Each case of change meant not only replacement (succession) but also the creation of a new management position or substantial changes in the position—as when one full-time chief of clinic was appointed to replace ten separate half-day clinic supervisors—or the addition of more support personnel, time, and budget to the position. In short, these clinic management changes can be viewed most productively not as simple management succession or the replacement of one manager by another, but as the institution of—or the increment of—management and supervisory activities in situations where there had been very little previously.

This interpretation of better performance following an increase in management and supervision is congruent with the cross-sectional findings of Bloom and Peterson [4] that coronary care units with full-time medical directors have better average performances than those with part-time directors or committee supervision. Those units under full-time medical directors perform better on a number of measures, including costs and patient survival and recovery rates.

The results from this variety of research designs, and sites and practitioners suggest that improvements in performances following study-instigated interventions and/or site-instigated management changes are not artifactual or irreproducible.

REFERENCES

1. Helfer, R. Estimating the quality of patient care in a pediatric emergency room. *Journal of Medical Education* 42:244-48, March 1967.
2. Becker, M. H. The Health Belief model and personal health behavior. *Health Education Monographs* 2:324-473, 1974.
3. Inui, T., et al. Improved outcomes in hypertension after physician tutorials. *Annals of Internal Medicine* 84:646-51, 1976.
4. Brook, R., and K. Williams. Evaluation of the New Mexico peer review system, 1971-1973. *Medical Care* 14: Supplement, December 1976.
5. Williamson, J., and M. Alexander. Continuing education and patient care research. *The Journal of the American Medical Association* 201:118-22, September 18, 1967.
6. Williamson, J. W. Formulating priorities for quality assurance activity: Description of a method and its application. *The Journal of the American Medical Association* 239(7):631-37, February 13, 1978.
7. Horn, S. D., and J. W. Williamson. Statistical methods for reliability and validity testing: An application to nominal group judgements in health care. *Medical Care* 15(11):922-28, November 1977.
8. Williamson, J. W., et al. Priority setting in quality assurance: Reliability of staff judgements in medical institutions. *Medical Care* 16(11):931-40, November 1980.
9. Williamson, J. W., H. R. Braswell, and S. D. Horn. Validity of medical staff judgements in establishing quality assurance priorities. *Medical Care* 17(4):331-46, April 1979.
10. Mann, F. D. Studying and Creating Change: A Means To Understanding Social Organization. In *Research in Industrial Human Relations*. Madison, WI: Industrial Relations Research Association, 1957.
11. National Training Laboratories Staff. Some Criteria for Useful Feedback. In *Reading Book, Laboratories in Human Relations Training*. Washington, D.C.: Institute for Applied Behavioral Science, 1968.
12. Payne, B. C., T. F. Lyons, et al. *The Quality of Medical Care: Evaluation and Improvement*. Chicago: Hospital Research and Educational Trust, 1976.
13. Neuhaus, E. R., T. F. Lyons, and B. C. Payne. Patient response to requests for permission to review medical records. *American Journal of Public Health* 66(11):1090-92, 1976.
14. Hare, R. L., and S. Barnoon. *Medical Care Appraisal and Quality Assurance in the Office Practice of Internal Medicine*. San Francisco: American Society of Internal Medicine, 1973.
15. Hulka, B. S., F. J. Romm, et al. *Physician Non-Adherence to Self-Formulated Process Criteria*. Chapel Hill, NC: University of North Carolina, Department of Epidemiology, 1978.
16. Hulka, B. S. and J. C. Cassel. The AAFP-UNC study of the organization, utilization, and assessment of primary medical care. *American Journal of Public Health* 63:494-501, 1973.
17. Neuhaus, E. N., T. F. Lyons, and B. C. Payne. Problems in case finding and data collection in ambulatory care settings. *American Journal of Public Health* 70:282-83, 1980.
18. White, N. H., et al. Ambulatory Care Quality Assurance Project.

(Unpublished manuscript). La Jolla, CA: Health Care Management Systems, 1976.
19. Hausser, D. L., P. A. Pecorella, and A. L. Wissler. *Survey-Guided Development II: A Manual for Consultants*. La Jolla, CA: University Associates, Inc., 1977.
20. Campbell, D. T., and J. C. Stanley. *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally, 1965.
21. Leavitt. H. J. Applied Organizational Change in Industry: Structural, Technological and Humanistic Approaches. In J. G. March (ed.). *Handbook of Organizations*. Chicago: Rand McNally, 1965.
22. Siegel, S. *Nonparametric Statistics*. New York: McGraw-Hill, 1956.
23. Guest, R. H. Managerial succession in complex organizations. *American Journal of Sociology* 68:47-56, 1962.
24. Guest, R. H. *Organizational Change*. Homewood, NJ: Irwin, 1962.
25. Torrence, E. P. Some Consequences of Power Differences on Decision Making in Permanent and Temporary Three-Man Groups. In A. P. Hare, E. F. Borgatta, and R. F. Bales. *Small Groups*. New York: Knopf, 1965.
26. Trow, D. B. Membership succession and team performance. *Human Relations* 13:259-69, 1960.
27. Trow, D. B. Executive succession in small companies. *Administrative Science Quarterly* 6:228-39, 1961.
28. Trow, D. B. Teamwork Under Turnover and Succession. Springfield, IL: National Technical Information Service, Technical Report 2, 1964.
29. Ziller, R. C., R. D. Behringer, and J. C. Goodchilds. Group creativity under conditions of success or failure and variations in group stability. *Journal of Applied Psychology* 46:43-49, 1962.
30. Blau, P. M. *The Organization of Academic Work*. New York: Wiley, 1973.
31. Burling, T., E. Lentz, and R. N. Wilson. *The Give and Take in Hospitals*. New York: Putnam, 1956.
32. Christensen, C. R. *Management Succession in Small and Growing Enterprises*. Boston: Harvard Graduate School of Business, Division of Research, 1953.
33. Eitzen, D. S., and N. R. Yetman. Managerial change, longevity, and organizational effectiveness. *Administrative Science Quarterly* 17:110-16, 1972.
34. Gouldner, R. W. The Problem of Succession in Bureaucracy. In K. Merton, P. Gray, B. Hockey, and C. Selvin. *Reader in Bureaucracy*. Glencoe, IL: Free Press, 1952.
35. Gouldner, R. W. *Patterns of Industrial Bureaucracy*. Glencoe, IL: Free Press, 1954.
36. Grusky, O. Role conflict in organization: A study of prison camp officials. *Administrative Science Quarterly* 3:452-72, 1959.
37. Grusky, O. Administrative succession in formal organizations. *Social Forces* 39:105-15, 1960.
38. Grusky, O. Managerial succession and organizational effectiveness. *American Journal of Sociology* 69:21-31, 1963.
39. Grusky, O. The effects of succession: A comparative study of military and business organization. In P. Janowitz. *The New Military*. New York: Russell Sage Foundation, 1964.

40. Heydebrand, W. V. *Hospital Bureaucracy*. New York: Dunellan, 1973.
41. Kahne, M. J. Suicides in mental hospitals: A study of the effects of personnel and patient turnover. *Journal of Health and Social Behavior* 9:255-66, 1968.
42. Mueller, E. H. The Relationship Between Teacher Turnover and Student Achievement. (Unpublished Ph.D. dissertation). University of Virginia, School of Education, 1969.
43. Revans, R. W. *Standards for Morale*. London: Oxford, 1964.
44. Bloom, B. S., and O. L. Peterson. End results, cost and productivity of coronary care units. *New England Journal of Medicine* 288:72-78, 1973.