The Validity of Nonexperimental Designs for Evaluating Health Services

O. L. DENISTON, MPH, and I. M. ROSENSTOCK, PhD

DOES EVALUATION of the effectiveness of health programs depend on random assignment of subjects to treatment and control groups? Some writers on evaluation state that this experimental design is the ideal, if not the only way to determine program effectiveness. But often it is impossible or unethical to use a true control group. At a time when many demonstration and ongoing programs in health services delivery are not properly evaluated because experimental design is not possible, it is imperative to develop alternative methods which still give valid results. Only in that way will it be possible to increase the rationality of program planning and consequently increase program effectiveness and efficiency.

Mr. Deniston is a research associate and lecturer and Dr. Rosenstock is a professor of public health administration, School of Public Health, University of Michigan. The investigation was supported by the National Center for Health Services Research and Development, Health Services and Mental Health Administration, Public Health Service grant HS 00386. Tearsheet requests to O. L. Deniston, School of Public Health, 1420 Washington Heights, Ann Arbor, Mich. 48104.

Recently, interest has grown in the suitability of nonexperimental design, frequently termed quasi-experimental designs, where true control groups cannot be used. These designs are intended for use in settings in which a program operator or evaluator has partial but not complete control over the situation. Thus, he may be able to control one or two but not all three of the following crucial elements in an evaluation: (*a*) which persons receive service, (*b*) when the service is to be provided, and (*c*) when evaluative measurements are to be made.

Campbell and Stanley (*1*) have been prominent among writers who propose several quasi-experimental designs which may help in evaluating social action programs when true experimentation is not feasible. One is the so-called interrupted time series design, in which a series of observations on some variable, or variables, of interest are made both before and after treatment. A comparison of the slopes and intercepts of the before and after series can be used to estimate the impact of the new treatment. This comparison was used, for example, to evaluate the effects of a statewide program in Connecticut to reduce traffic deaths by arresting speeders (*2*). Within that design, further evidence was sought to confirm the plausibility of a cause-effect relationship, for example, the proportion of traffic violations that were for speeding compared with other kinds of violations.

In making such supplementary investigations, Campbell (*2*) was using an analysis which Suchman (*3*) terms "intervening variable analysis" and which we term "internal analysis." Kelman and Elinson (*4*) and Heyman (*5*) also describe this approach to attributing causality. One application of this approach is to examine the experiences of several groups within the program, for example, a group that has received the full range of intended services can be compared with groups that for various accidental reasons have failed to receive one, two, or several intended services. By comparing outcomes among the several groups, one can presumably estimate the importance of each of the services provided.

Campbell and Stanley also recommend, when nothing better is available, the use of the nonequivalent control group (or control series) design in which assignments to treatment and control conditions are not random; rather the groups are natural collectivities deemed but not proved to be similar. Other recommended designs represent variations on these three themes: (*a*) the time series, (*b*) the intervening variable or internal analysis, or (*c*) the nonequivalent control group (or control series design).

Another design which Campbell and Stanley (*1*) term nonexperimental rather than quasi-experimental is the single group "before-after" design: the status of an outcome (dependent) variable is measured in a single group both before and after the introduction of some treatment. Although research specialists have long pointed out the dangers in drawing conclusions from this design, it is widely used in public health programs.

A review was made of papers published in the American Journal of Public Health and HSMHA Health Reports (now Health Services Reports) for 1970–71 to identify health program evaluations. In all, evaluation was reported in 40 studies. Only five of the 40 used a true control group. At the opposite extreme, seven used testimonials, that is, unsolicited comments from self-selected consumers. Ten used a nonequivalent comparison group, and four used internal or intervening variable analysis. Fourteen, or 35 percent of all reported studies, used a single treated group measured before and after treatment. Because the before-after design is so popular, it is worth including in studies of evaluation methods to determine whether it may at times provide reasonable estimates of program effectiveness.

Campbell and Stanley argue convincingly that the alternatives to true experimentation pose a number of problems in analysis and interpretation, yet they are recommended where nothing better is possible (*1*).

In the present study we attempt to determine the validity of various approaches to measuring effectiveness. But to estimate the validity of nonexperimental design, we need settings where true control groups do exist. The research strategy used in this paper was to estimate program effectiveness in a number of settings by each applicable nonexperimental design; we then assessed the validity of each design in that setting by comparing the estimates of program effect with a criterion estimate provided by a true randomly assigned control group.

Program Description

The program used was the Michigan Arthritis Control Program, a cooperative venture of the University of Michigan Medical School, Ann Arbor Veterans Administration Hospital, Wayne County (Michigan) General Hospital, and Henry Ford Hospital (Detroit). (Unpublished paper by I. F. Duff, professor of internal medicine, University of Michigan, and co-workers entitled, "Comprehensive Care of Patients with Rheumatoid Arthritis: Some Results of the Regional Arthritis Control Program in Michigan.") The objective of the program was to prevent, reduce, or delay the development of disability and deformity.

All qualified self-referred patients and patients referred by a physician who appeared at the clinics during the study period were enrolled in the program. The patients must not have had primary care for their arthritis at any of the participating institutions. These patients also had to meet the criteria of (*a*) a diagnosis of definite or classical rheumatoid arthritis of less than 7 years' duration with onset after age 16, (*b*) plans to remain in the area and the ability and willingness to follow a study protocol, (*c*) a signed agreement of willingness to participate in the experiment, after the study aims and methods were described to them. None of the patients who met the preceding criteria declined to participate. After being classified by sex and duration of disease, patients were randomly assigned to either the treatment or experimental group. The treatment group received comprehensive care, and the control group received conventional care. Both groups received the standard medical care which is offered to all patients with arthritis in the institutions participating in the program.

The two treatment programs differed in several dimensions, the basic difference involving accessibility to and utilization of professional manpower to meet defined patient needs. Every patient assigned to the comprehensive program was referred to the occupational therapist, physical therapist, social worker, and visiting nurse, while, in the control group, judgment regarding the necessity of referrals was handled in the customary way by attending clinic physicians. In contrast with the control group, the treatment of the comprehensive group also included the following elements.

Conference presentation. At least four times in the first year and other times as required, the patients' treatment was discussed by the treatment team, which included a rheumatologist, an occupational therapist, a physical therapist, a social worker, a visiting nurse, and counselors and other health professionals. The conference was designed to define the patient's needs, spell out reasonable objectives, and develop specific recommendations for attaining those objectives.

Continual monitoring of patients and their progress. Progress of patients was monitored by a review of clinic visits and by home visits made by the visiting nurse, often accompanied by the physical therapist and occupational therapist. Progress reports were made to the treatment team and necessary steps were taken to correct any inappropriate situations; when necessary, the patient's problems were reconsidered in the conference setting.

The original experimental plan called for a 5-year study of at least 500 patients. The program was funded by a Federal contract on July 1, 1969, but on September 11, 1969, 3 days after the first patient was enrolled, the investigator was notified that the contract would be terminated after 1 year. Based on hopes of changing that decision as well as a desire to accomplish as much as possible, the program was implemented as scheduled. Because new funding was not obtained, the last patient was enrolled on April 30, 1970, and the final assessment was made in November 1970. Because it was believed necessary to follow patients 5 years to make an adequate study, the premature termination of the program makes it unrealistic to expect a clear-cut test of the program hypotheses. The program did, however, provide some data which allow discussion of alternative evaluation methodologies.

Description of patients studied. Of the 80 patients enrolled in the program, 39 were assigned to the comprehensive treatment (experimental) group, and 41 were assigned to the conventional treatment (control) group. The sex and age of patients studied are given in table 1.

Males in the study were somewhat older than the females and were considerably overrepresented compared with national estimates which indicate that females have rheumatoid arthritis three times as often as males.

Patients were enrolled over a 9-month period. Because care and assessment for study purposes had to terminate in September 1970, patients were enrolled for different periods of time. In the experimental group, all 39 patients were reassessed after 4 months. Twenty-four of these also had a second reassessment at 8 months while eight had a third reassessment at 12 months. Thirty-six patients in the control group were reassessed after 4 months; 23 of these had the second reassessment after 8 months, while 10 had the third reassessment at 12 months. Two patients in the control group died before the first reassessment at 4 months, and three additional control patients did not keep their first reassessment appointment although they remained in the study and were assessed after 8 months.

Evaluation approaches. The program was a true experiment where a comparison of change, or improvement scores between the two groups of arthritic patients would have yielded definitive estimates of program effectiveness if the experi-

Table 1. Sex and age of study patients

Group	Number	Percent of group	Average age
Comprehensive:			
Male	17	44	48.6
Female	22	56	41.0
Total	39	100	44.3
Conventional:			
Male	15	37	46.1
Female	26	63	44.1
Total	41	100	44.8
Total:			
Male	32	40	47.4
Female	48	60	42.7
Total	80	100	44.6

ment had been completed. But the purpose of the present evaluation study was to use alternative techniques for estimating program impact. Our task, thus, involved two stages; first, to devise alternative evaluation approaches to the classical experiment and, second, to determine the validity of each alternative method, using the true control group as the criterion.

The following indices, or dependent variables, representing disease activity, were used to assess program effect:

Duration of morning stiffness (minutes)
Right hand grip strength (mm mercury)
Number of involved joints (objective joint evaluation)
Most troublesome joints (clinical judgment)
Erythrocyte sedimentation rate (Westergren method)
American Rheumatism Association functional classification (clinical judgment on a 4-point scale).

Objective evidence of joint involvement was expressed in terms of tenderness or pain on motion, swelling, heat, redness, limitation of motion, and deformity. Joints having one or more of these characteristics on clinical examination were classified as "involved joints."

"Troublesome joints" were identified by the examining physician on the basis of objective evidence and the patient's complaints.

Right hand grip strength was measured by the patient's ability to compress a folded, inflated sphygmomanometer cuff under standardized conditions.

Before proceeding to a consideration of nonexperimental designs used to estimate program effect, the true impact of the program is briefly summarized as follows. As indicated earlier, definitive tests of the program hypotheses were precluded by the premature termination of the program; nevertheless, the data generally suggest superiority of comprehensive treatment over conventional treatment. On each of the measures studied, greater improvement was exhibited in the comprehensively treated group than in the conventionally treated group.

Although no single difference was statistically significant, probably because of the small numbers and short treatment duration, as duration of treatment increased, comprehensive treatment became consistently more effective. Thus, over the short duration of this study of a relatively small sample, a small but systematic superiority was shown in those patients randomly assigned to comprehensive treatment groups over those randomly assigned to conventional treatment groups. It is not unreasonable to expect that the superiority of comprehensive treatment would have become even more clear cut had the program continued as planned.

The major question considered in this paper was: How closely could true program effect be assessed had a true control group not been available?

Before-after without control group. The first nonexperimental method considered was the traditional "before-after" design (sometimes termed pretest, post-test) without a control group. The basic assumption is that patients would not have changed had they not entered the experimental group. This method, although severely criticized by methodologists such as Campbell and Stanley (*1*), is, as indicated earlier, the most frequently used method in the evaluation of health service programs as reported in the public health literature.

The results from this method are reported in table 2. The "before" scores for all patients in the

Table 2. Changes during enrollment in Arthritis Control Program, experimental group, by length of enrollment

Index	Changes from before-after							
	Before (N=39)		4 months (N=39)		8 months (N=24)		12 months (N=8)	
	Average	Median	Average	Median	Average	Median	Average	Median
Duration of morning stiffness	26.1	....	+ 2.8	....	− 0.3	....	[1] +10.0	....
Grip strength	142.0	....	[1] +22.6	....	[1] +43.0	....	+31.2	....
Number of involved joints	30.7	29.8	[1] + 4.8	[1] +7.8	[1] +10.3	[1] +11.5	[1] +15.5	[1] +20.5
Number of troublesome joints	11.7	9.0	[1] + 2.2	[1] +1.1	[1] + 3.9	[1] + 3.0	+ 4.6	+ 3.0
Sedimentation rate	48.3	....	+ 5.3	....	[1] +12.3	....	[1] +25.0	....
Functional classification	1.82	1.94	− .09	− .03	+ .28	+ .25	+ .20	+ .25

[1] Indicates statistical significance at or below the 0.05 level.

Table 3. Slope of regression lines for duration and severity of disease prior to treatment and differences between predicted and actual disease scores following treatment

Index	Regression model		Difference between prediction and after scores		
	beta	P[1]	4 months	8 months	12 months
Duration of morning stiffness	+0.30	0.04	+ 4.0	+ 2.1	+13.6
Grip strength	− .26	.06	+24.0	+45.0	+34.3
Number of involved joints	+ .12	.04	+ 5.3	+11.3	+16.9
Number of troublesome joints	+ .04	.27	+ 2.4	+ 4.2	+ 5.1
Sedimentation rate	+ .26	.02	+ 6.3	+14.3	+28.1
Functional classification	+ .01	.01	− .05	+ .32	+ .32

[1] Probability of observing a "b" of this magnitude or greater, if true b were zero.

experimental program are reported as baseline data with which reported changes can be compared. The scores are presented so that a plus (+) indicates improvement; a reduction is an improvement for all indices except for grip strength where an increase is an improvement.

There is a question concerning the appropriateness of using parametric statistics for analysis of the data concerning three of the indices; that is, numbers of involved joints, number of troublesome joints, and functional classification. The Cooperating Clinics Committee of the American Rheumatism Association uses medians rather than means in such analyses and does not perform statistical tests of differences (*6*). We compared the means and medians for these three indices for each of the treatment periods and examined the differences from before to after treatment using both Student's t test and the Wilcoxon matched-pairs signed-ranks test. It can be seen in table 2 that means and medians are quite similar. In each of the nine comparisons, the results of the t test and the Wilcoxon test were identical. Because the different procedures yield similar results, only means and t tests are reported in subsequent tables for ease of presentation.

No patient who entered the program left the program (except for two persons in the control group who died before the 4-month assessment), and the duration of exposure (4, 8, or 12 months) was determined entirely by date of entry; no patient chose the duration of his treatment. Inspection of the data for patients with different treatment periods indicates that those with 8 months of treatment showed more improvement after 8 months than after 4, and those with 12 months of treatment showed more improvement after 12 months than after 8 and after 8 months than after 4.

Before-after, without control group, weighted by progression of disease. In a second, similar approach, we attempted to account for the often expressed opinion that rheumatoid arthritis is a progressive disease and that the condition of the patient without medical treatment would worsen over time rather than remain constant as is assumed in the before-after approach.

The first step was to determine whether there was an association between duration of disease at time of entry into the program and the indices or dependent variables being used. The data used were observations of all 80 patients in the program at the time of initial assessment. Although patients were accepted in the program who had had the disease as long as 7 years, there was considerable variation in duration among patients. The average duration of the disease at first visit was 2.3 years and the range was from 2 months to 83 months. A linear regression model was used to estimate the relationship between duration of disease at first visit and the dependent variables. For each of the six indices, the slope of the regression line supported the belief that arthritis increases in activity with increased duration. For five of the six indices, the slope was significantly different than 0.

To estimate what experience the patient would have had if he had not participated in the program, individual scores were computed by substituting in the prediction formula: $y = a + bx$
where

a = score at entry

b = slope of the least squares regression line

x = interval in months between entry and final assessment (that is, 4, 8, or 12 months).

The beta (b) values and the estimates of program effect using this model are reported in table 3. In each instance, this procedure results in a predic-

Table 4. Comparison of patients in experimental and drug studies at entry into study

Patients	Sex		Average duration (months)	Average age at onset	Morning stiffness (minutes)	Grip strength (mm mercury)	Number of involved joints	Sedimentation rate	Functional classification
	Male	Female							
Experimental	17	22	25.5	44.4	26.1	142.0	30.7	48.3	1.82
Drug study	8	6	42	49.5	131.2	122.6	34.7	52.4	2.14
Drug study (adjusted)[1]	8	6	25.5	44.4	125.1	122.1	32.1	45.0	1.93

[1] Scores were adjusted for duration of disease and age at onset using regression analysis *b* values obtained from study of Arthritis Control Program patients at entry into the study.

tion that the health of patients would have deteriorated (higher scores on all indices except grip strength) rather than have stayed the same. We then compared the predicted and actual scores to estimate program effect. With the adjustment for disease duration, the apparent program impact is even greater than for the simple before-after design.

Before-after with nonrandomly assigned comparison groups. The next approach was to use some available comparison groups; a group of patients similar to those enrolled in the program but whose care was expected to be similar to the control group. Two somewhat different groups were used.

Before proceeding, it should be pointed out that the two groups selected for convenience were by no means ideal. The first, a group of patients participating in a study of the relative effectiveness of a new drug, differed from those in comprehensive treatment in the control program in several important ways that will be described later. The second group, drawn from Tecumseh, Mich., consisted of only four patients, a group far too small to serve as a basis for definitive comparison. Nevertheless, in the context of the present study, they were used because our aim is to throw light on the kinds of conclusions that might be drawn from the use of nonrandomly assigned groups and, therefore, to raise questions about conditions under which such groups may or may not be legitimately used.

Drug Trial Sample

Soon after the Arthritis Control Program ended, the program director participated in the trial of an anti-inflammatory, nonsteroidal experimental drug. That study, spanning 16 weeks, used a crossover design in which patients received the experimental drug for 6 weeks and aspirin for 6 weeks, with the order randomized and a placebo separating the two treatments. Evaluation of the effects of the experimental drug has not been completed. Except for the use of the new drug, the kind of medical care received by the patients was similar to that received by the true control group in the Arthritis Control Program. The frequency of medical assessments, however, was far different. Patients using the trial drug were examined each week and were interviewed daily by telephone.

Of the 20 patients admitted to this study, 14 met the criteria given for the Arthritis Control Program of diagnosis and duration of rheumatoid arthritis while the other six met the diagnostic but not the duration criteria. In table 4, comparisons are made of the group of 14 similar patients and all patients in the experimental group at entry. (The index, troublesome joints, was not included, because it was defined differently in the two studies.)

There were considerable differences between the two groups. For example, there was a higher proportion of men in the drug study, patients in the drug study were slightly older at onset, had had the disease almost twice as long, and had scores on all indices indicating greater disease activity.

It was reported earlier that a positive relationship was observed between duration of disease and disease activity among patients in the study at entry into the program. We also examined the relationship between age at onset and disease activity. Here we found that for four of the five indices, the older the patient was at onset, the greater the disease activity. For the index, grip strength, the older the patient was at onset, the better his grip strength. These relationships were not as strong as those between duration at entry and disease activity; only the relationship of the sedimentation rate was statistically significant at the 0.05 level.

In an attempt to standardize the two groups by adjusting for initial differences between them, the results from the regression analysis were applied to the scores of the drug trial comparison group in the manner described earlier. Because the regression values were similar for males and females, no adjustment was attempted for the slight difference in sex composition of the two groups. The adjusted scores of the comparison group were closer to the scores of the experimental group in all patients, but the groups still differed significantly in duration of morning stiffness. This difference is not because of a few extreme scores in either of the groups. Only six of the 14 patients in the comparison group were experiencing 1 hour or less of morning stiffness at entry while none of the patients in the experimental group experienced as much as 1 hour; only five patients in this group reported more than 30 minutes of stiffness.

To estimate program effectiveness using this comparison group, we looked (*a*) at absolute change between scores at entry and following 4 months of treatment and (*b*) change as a proportion of possible change (table 5).

This analysis suggests that the Michigan Arthritis Control Program was not nearly as effective as had been suggested by the before-after design; in fact, the program might be judged slightly less effective, and certainly no more effective than the comparison treatment; of course, the effects of the new drug represent a potential confounding factor. There were small differences on four of the five indices, two favoring the experimental group and two favoring the comparison group. On the fifth index, duration of morning stiffness, the results are doubtful because they were greatly influenced by three patients in the drug study, whose post-treatment response of "all day" was not probed. None of the patients reported "all day" at entry into the study.

These results differ greatly from earlier results based on the before-after design and the progression of disease design. Without benefit of the true control group, the observed changes might suggest that both treatments were effective, the drug trial treatment perhaps more effective in increasing grip strength and reducing the number of involved joints. Two possible explanations are that for some patients the new drug was more effective than the standard drugs used for patients in the experimental group or that the increased attention associated with weekly assessments and daily interviews were responsible for improvement. In any

Table 5. Change scores for experimental patients and 14 "similar patients"

Index	Experimental (N=39)		Comparison group (N=14)		Relative improvement in two groups	
	Absolute improvement	Percent improvement	Absolute improvement	Percent improvement	Absolute improvement	Percent improvement
Duration of morning stiffness [1]	2.8	10.7	−105.7	−80.9	+108.5	+91.6
Grip strength	22.6	16.2	29.4	23.9	− 6.8	− 7.7
Number of involved joints	4.8	15.6	6.5	18.7	− 1.7	− 3.1
Sedimentation rate	5.3	11.0	4.2	8.0	+ 1.9	+ 3.0
Functional classification	− .09	− 2.7	− .07	− 3.3	+ .02	+ .6

[1] For 3 drug patients who reported morning stiffness as lasting "all day," the duration was arbitrarily considered to be 12 hours.

NOTE: + indicates greater improvement in experimental patients; − indicates greater improvement in comparison patients.

Table 6. Comparison of experiences of patients from Tecumseh sample and experimental group

Group	Number	Average age at onset	Average duration at time of observation (months)	Morning stiffness	Grip strength	Involved joints	Troublesome joints	Sedimentation rate	Functional classification
Tecumseh group:									
Total	17	44.1	111.4	154.7	169.6	20.3	2.2	27.5	1.75
Meeting criteria	4	45.2	43.0	60.0	187.0	28.8	1.25	20.0	1.75
Experimental:									
12 months treatment	8	44.2	39.0	13.6	174.0	17.0	5.0	31.0	1.63
8 months treatment	24	48.6	36.3	30.6	175.1	21.7	8.9	37.9	1.76
4 months treatment	39	44.4	29.5	23.3	164.7	25.9	9.5	43.0	1.91

event, if the drug trial group was believed to approximate a true control group, comprehensive care would not appear to be consistently more effective.

Tecumseh sample. A second comparison group was assembled from residents of Tecumseh, Mich., who were participating in a community health study. It was initially believed that data would be easily obtained for this group since their health has been under continuous study since the late 1950s (7). Patients suspected of having rheumatoid arthritis were identified from the records, and during the summer and fall of 1971 clinical examinations were made to obtain adequate diagnosis and measures of the indices.

Initial discussion with knowledgeable persons suggested there might be 30 to 40 people in the community who met the diagnostic and duration criteria of the Arthritis Control Program. This number was deemed adequate for use as a comparison group.

The results, however, were discouraging because, of 142 examinations made, a total of only 17 persons with probable or definite rheumatoid arthritis were found; 10 of these diagnoses were definite, and only four of the 10 were of less than 7 years' duration. Average scores for the two groups of these patients are reported in table 6; the comparison with the experimental group is made on the group meeting both diagnostic and duration criteria, although the group includes only four patients. We are more concerned with the kinds of inferences that might be drawn from the use of such a comparison group than with the stability of scores based on our four patients.

The analytic approach is different in this comparison from the drug study group because no change scores are available for patients in the Tecumseh study. We therefore compared "after only" scores, whose usefulness depends on the validity of two questionable assumptions: (*a*) that the Tecumseh and experimental groups were similar at the beginning of the Arthritis Control Program and (*b*) that the patients in the experimental group would have received care similar to the patients in the Tecumseh group had they not been in the experimental group. Better scores for the patients in the experimental group would indicate the program's effect.

The index "troublesome joints" is also "troublesome" in this comparison although the forms and definitions were identical to those used in the Arthritis Control Program. When the size of the difference between the two groups reported in table 6 was observed, doubts were expressed as to whether this was a true difference or whether there might be some other explanation. It seems there may have been some unreliability in results caused by differing diagnostic styles of examining physicians. The one physician who made most of the examinations in the Tecumseh study made only a few of the Arthritis Control Program examinations, but in those he reported many fewer troublesome joints than the program average (2.67 as compared with 11.35 for all other physicians). He did not differ on the related index, total number of involved joints (31 compared with 30.7), suggesting some caution in interpreting the difference between groups for the troublesome joint index.

The experience of the group of four patients from Tecumseh is similar to that of the group of eight patients in the experimental group in terms of age at onset and duration of disease at time of observation. The Tecumseh patients experienced considerably more morning stiffness, exhibited slightly greater grip strength, had more involved joints, fewer troublesome joints, lower (in fact "normal") sedimentation rates, and the functional classifications were slightly lower. Overall, it might be reasonable to conclude that the groups do not differ greatly, indicating the Arthritis Control Program had no great impact.

The data in table 6 raised doubts about the assumptions upon which the design was based, specifically, the assumption that the Tecumseh and experimental groups were similar at the beginning of the Arthritis Control Program. That assumption would imply that during the period of the Arthritis Control Program the Tecumseh patients would have had to experience as great improvement as the experimental group even though they were not enrolled in any special experimental program. That implication seems highly unlikely.

Additional observations, already reported, lend further support to the view that the experimental and Tecumseh groups were not similar at the beginning of the Arthritis Control Program. Data from tables 4 and 5 showed that: (*a*) even though the beginning of the experimental and drug trial programs were separated in time by more than a year, patients in both groups were similar at the time they entered their respective programs and (*b*) each of these groups showed considerable improvement during the course of their enrollment.

It will be recalled that one criterion for participation in the Arthritis Control Program was that the patient be new to the hospital. That criterion was not used in the drug study. This difference allowed a comparison of six patients in the drug study who were new to the hospital with the eight who were already undergoing treatment. On four of the five indices of disease activity the new patients in the drug study indicated greater disease activity at entry than did the patients already undergoing treatment.

Consideration of all available data suggests the following picture. Rheumatoid arthritis is a disease that varies greatly among different persons and with groups of people over time. Although the disease seems to be generally progressive, there will be "flares" and "remissions" within individual persons. It is suggested, both from the comparison of new and old drug study patients and the discussion of the results from the Tecumseh group, that a tendency existed for patients with rheumatoid arthritis to seek care from or to be referred to a medical center at an unusually severe stage of their disease.

One might thus consider the population of patients with rheumatoid arthritis as representing all stages of disease activity with those experiencing unusually severe disease activity tending to seek some treatment that will be more effective than the treatment available to them under ordinary circumstances. Such people will thus be disproportionately represented in a group seeking care from a medical center or referred to a medical center by a family physician. If this argument is correct, then the improvement over time in both experimental and drug study groups could represent the effects of statistical regression toward the mean as well as the effects of treatment. That is, a group selected for its extreme score on some dependent variable may well show apparent improvement over time that is attributable to an artifact in their selection. From data presented thus far we have no way of estimating the relative amount of improvement caused by each factor from data obtained from the nonexperimental designs.

In summary thus far, in design one ("before-after" without control group) the results suggested that the experimental program had a substantial effect on the patients with a tendency for greater improvement to be associated with longer enrollment in the project. When the scores for disease activity of the experimental group were adjusted by disease duration at time of entry, the apparent program impact was even greater.

Relative improvement of the experimental group was less in design two when they were compared with a nonrandomly assigned comparison group. And when compared with a second, nonrandomly assigned comparison group, the status of the patients' health in the experimental group was not superior after treatment.

Finally, it has been argued that patients in the experimental group may have been experiencing a period of greater disease activity at the time they entered the study than did a small sample of the general population of arthritics studied a year later.

Table 7. Changes in severity of disease states in patients assigned to experimental and control groups

Index	4 months' treatment			8 months' treatment			12 months' treatment		
	Experimental (N=39)	Control (N=36)	Difference [1]	Experimental (N=24)	Control (N=23)	Difference [1]	Experimental (N=8)	Control (N=10)	Difference [2]
Duration of morning stiffness	+ 2.8	− 0.9	+3.7	− 0.3	+ 0.7	− 1.0	+10.0	+ 4.4	+ 5.6
Grip strength	+22.6	+22.6	0	+43.0	+12.4	+30.6	+31.2	+27.3	+ 3.9
Number of involved joints	+ 4.8	+ 5.6	− .8	+10.3	+ 5.5	+ 4.8	+15.5	+11.7	+ 3.8
Number of troublesome joints	+ 2.2	+ 2.1	+ .1	+ 3.9	+ 4.9	− 1.0	+ 4.6	+ 1.7	+ 2.9
Sedimentation rate	+ 5.3	+11.4	−6.1	+12.3	+ 9.9	+ 2.4	+25.0	+ 6.5	+18.5
Functional classification	− .09	− .05	− .04	+ .28	+ .04	+ .24	+ .2	+ .1	+ .1

[1] Based on the experimental group, that is, + indicates more improvement for the experimental group; − indicates more improvement for the control group.

[2] The probability of more favorable outcomes occurring in the experimental group on all 6 measures is 0.016 if there were no true differences between experimental and control groups.

Comparison of experimental group and true control groups. In this comparison, as throughout the preceding discussion, conclusions about treatment efficacy are limited by small numbers and the brevity of the periods of treatment. Data, however, are more adequate for estimating the validity of the foregoing quasi-experimental designs.

Although there were slight differences in average scores on the various indices between the patients in the experimental and those in the control groups at time of entry, they were not significant with one exception to be noted later. We will therefore present only the changes that occurred in patients with varying durations of treatment.

None of the differences reported in table 7 between experimental and control groups are statistically significant. A trend is evident that as duration of treatment is increased, however, the systematic superiority of comprehensive treatment becomes more clear cut, although it remains, small. At 4 months there is virtually no systematic variation between the two groups, by 8 months comprehensive treatment seems to yield somewhat better results, and by 12 months the comprehensive group is superior on all six indices. This difference between the two groups is statistically significant ($P = 0.016$). One wonders whether the superiority of comprehensive treatment would have become even more striking had the experiment continued for the full 5 years with a larger group of patients.

Although the comprehensive program is thus shown to be slightly more effective than the conventional, public health practitioners would probably conclude that both forms of treatment are effective with a slight systematic advantage for comprehensive care over the short run (1 year) which might become greater if a longer treatment span were provided. A question is raised, however, concerning the extent to which improvement may represent effects of regression resulting from having served a population with rheumatoid arthritis who sought care at an unusually severe stage of disease.

Data from previous tables are summarized in a composite given in table 8, and show how closely the experience of the true control group is estimated.

Examination of the data in table 8 suggests that the status of each of the two comparison groups (drug study and Tecumseh patients) is not substantially or systematically different from that of the control group (except for duration of morning stiffness and number of troublesome joints which, as previously indicated, may reflect differences in reporting and diagnosing rather than differences in activity of disease). On some measures, the status of the control group is better; on others, each comparison group appears better.

Earlier discussion led to the same conclusion regarding patients in the experimental group, that is, compared with patients in each comparison group, the experimental group was not substantially or systematically different. Therefore, the systematic, though slight superiority of experimental care, revealed by comparison with the true randomly assigned control group was not closely estimated by the quasi-experimental designs.

Validity of Quasi-Experimental Designs

Before-after design. It is clear that the "before-after" model without controls, and a related approach based on "before-after" with adjustments for the progressive nature of the

Table 8. Comparison of control group experience with drug study

Group	Duration of morning stiffness			Grip strength			Number of involved joints			Number of troublesome joints		
	Before	After	Difference[1]	Before	After	Difference[1]	Before	After	Difference[1]	Before	After	Difference[1]
Drug study	131.2	263.9	−105.7	122.6	152.0	+29.4	34.7	28.2	+ 6.5			
Control:												
4 months	17.6	18.5	− .9	134.6	157.2	+22.6	29.0	23.4	+ 5.6	10.3	8.1	+2.1
8 months	17.5	16.8	+ .7	136.1	149.4	+13.3	29.0	23.5	+ 5.5	12.6	7.7	+4.9
12 months	19.1	14.7	+ 4.4	156.9	184.2	+27.3	27.7	16.0	+11.7	11.9	10.2	+1.7
Tecumseh		60			187			28.8			1.25	

[1] + indicates improvement in health status over time; − indicates deterioration in health status over time.

disease are clearly invalid in the present setting. One source of invalidity may be due to the tendency of patients to enter the experimental program during an unusually acute stage of their disease.

The point cannot be emphasized too strongly that public health programs today are generally being evaluated on a before-after basis with no attention given to the probability of regression even where the study group is selected because it is an extreme group on certain important dependent variables where regression effects may well account for much of the outcome that is usually attributed to the program effort.

For example, in a traditional food protection program, routine inspections are made at random and observed violations are called to the operator's attention. If evaluations of program success are based on whether those particular violations are corrected at a subsequent visit, it may well be that regression may account for some of the usually observed correction. If evaluation were based on the total number of violations observed at a point in time, however, regression should not be a problem if inspections were made at random and at unannounced times.

Inspections, however, might be based on complaints. In this instance, it is likely that the establishment is in an "extreme" condition at the time of the complaint. An inspection made quickly after receipt of such a complaint may find many violations, although later followup inspections would find fewer. An apparent improvement may well be caused by regression instead of a positive change.

Many local health departments now deliberately wait a few days before responding to nuisance complaints. This procedure has evolved from experience which showed the alleged condition was noted much less frequently than when complaints were processed promptly.

and Tecumseh patient experiences

Sedimentation rate			Functional classification		
Before	After	Difference[1]	Before	After	Difference[1]
52.4	48.2	+ 4.2	2.14	2.21	−0.07
41.9	30.5	+11.4	1.79	1.84	− .05
34.5	24.6	+ 9.9	1.83	1.79	+ .04
30.4	23.9	+ 6.5	1.88	1.77	+ .11
......	20.0	....	....	1.75	....

To illustrate the problem with another example, consider a large group of drivers who had had six or more automobile accidents during the preceding year and who were consequently required to participate in a safe driving course. In all probability the average accident rate for that group during a subsequent year would drop regardless of the effects of the program. Moreover, in a large group of drivers who had had no accidents in the preceding year, the rate would almost certainly increase. If they were given the same educational program, one might erroneously conclude that education is effective for poor drivers but deleterious to good drivers.

Thus, the problem of regression may often go unrecognized in both personal health and environmental health programs. We believe that where clients are not selected because they are extreme on some dependent variable, regression effects may be less important. Future studies will permit more adequate description of settings in which before-after and related designs can be relied upon with greater confidence.

Comparison group design. Whereas the before-after approach overestimated program effect, the first comparison group underestimated program effect, that is, its performance was similar to the experimental group. The extent to which the drug study group performed relatively well because of the efficacy of the new drug in some patients, or because they were given greater attention, or because of some combination of these or other factors cannot be determined, though some impact of the combination of drug efficacy and attention seems probable.

We do not believe the patients in the second comparison group were similar. The Tecumseh arthritics, though small in number, represented a cross section of patients at all stages of disease while the experimental patients were probably a largely self-selected group who sought care at a hospital clinic during an atypical stage of their disease. We believe that over the 12-month course of study they approached the status of the Tecumseh group as a function both of regression and efficacy of treatment.

Conclusions. In this, the first of a series of papers concerning the validity of nonexperimental or quasi-experimental designs in estimating program effectiveness, the use of two, single group before-after designs without controls, and the use of two designs using nonequivalent control groups were shown to yield invalid estimates of program effectiveness when compared with the experiences of a true control group.

The single-group before-after designs overestimated program effectiveness while the nonequivalent comparison groups underestimated program effectiveness.

It was argued that because the treatment group was a self-selected group of rheumatoid arthritics randomly assigned to two treatments, one could expect regression effects to result in an apparent reduction of activity of the disease even if no treatment had been provided. In such patients, a single-group before-after design cannot separate regression from treatment effects. Nor can arthritics in the general population serve as a valid control in such a setting since most of them would probably not be in an atypical state.

A comparison group of hospital outpatients, about half of whom were new to the clinic, did not provide valid indices of program effectiveness, probably because this group received both a new drug and more attention than the patients in the experimental program.

Comparisons between experimental and true (randomly assigned) controls yielded results that support the regression hypothesis. Both groups showed progressive improvement with a tendency for the experimental (comprehensive) care to become more consistently superior to control (conventional) care, although the margin of superiority in this short-run study was small.

The present study implies that whenever groups are selected because they show extreme scores on some health condition whose severity changes periodically, the interpretation of any intervention must be tempered by the tendency for extreme scores to regress toward the mean over time. Additional studies will be needed, and are being made, to determine the conditions under which "imperfect" designs may or may not safely be used.

REFERENCES

(*1*) Campbell, D. T., and Stanley, J. C.: Experimental and quasi-experimental designs for research. Handbook of Research on Teaching, Rand McNally & Co., Chicago, Ill., 1963.

(*2*) Campbell, D. T.: Reforms as experiments. Am Psychol 24: 409–429, April 1969.

(*3*) Suchman, E. A.: Evaluation research. Russell Sage Foundation, New York, 1967.

(*4*) Kelman, H. R., and Elinson, J.: Strategy and tactics of evaluating a large-scale medical care program. Med Care 7: 79–85, March–April 1969.

(*5*) Heyman, M. M.: Criteria and guidelines for the evaluation of in-service training. U.S. Government Printing Office, Washington, D.C., 1968.

(*6*) Cooperating Clinics Committee of the American Rheumatism Association: A controlled trial of cyclophosphamide in rheumatoid arthritis. New Engl J Med 283: 883–889, Oct. 22, 1970.

(*7*) Epstein, F. H., et al.: The Tecumseh study: design, progress and perspectives. Arch Environ Health 21: 402–407 (1970).