STATISTICS IN PRACTICE

CROSSOVER AND SELF-CONTROLLED DESIGNS IN CLINICAL RESEARCH

Thomas A. Louis, Ph.D., Philip W. Lavori, Ph.D., John C. Bailar III, M.D., Ph.D., and Marcia Polansky, M.S.

Abstract Crossover studies (clinical trials in which each patient receives two or more treatments in sequence) and self-controlled studies (in which each patient serves as his or her own control) can produce results that are statistically and clinically valid with far fewer patients than would otherwise be required.

We investigated the use of the crossover design in the 13 crossover studies that appeared in the *Journal* during 1978 and 1979. We considered the following important features of design and analysis as they applied to these studies: the method by which patients were assigned to initial treatment (only 7 of 13 studies used random assignment); the determination of when to switch treatments (10 of the 13 used a time-dependent rule, and 3 a less appropriate disease-state–dependent rule); blinding of the crossover point (in only 3 of the 13 studies was the crossover point concealed, but in 4 of the remaining 10 concealment was impossible); assessment of the effects of the order of treatments (included in only 1 of the 13 studies); and the use of at least minimally acceptable statistical analysis (11 of the 13 studies had such an analysis).

We also report briefly on 28 additional studies of a single treatment each, in which each patient served as his or her own control before or after treatment or both. The scientific issues were much the same as in crossover studies except that self-controlled comparisons of treatments tended to be less precisely designed and conducted and to focus on clinical problems and patient groups that are especially difficult to study. (N Engl J Med 1984; 310:24-31.)

An earlier paper in this series[1] discussed issues bearing on the strength of scientific inferences from parallel comparisons of clinical treatment — studies in which each patient receives only one treatment and responses in one group are compared with those in another. This report extends the discussion to crossover studies — those in which each patient receives two or more treatments in sequence and outcomes in the same patient are contrasted. We also comment briefly on self-controlled studies, in which a single treatment under study is evaluated by comparison of patient status before and after treatment. We consider the role of the crossover design in clinical research, using 13 crossover studies from Volumes 298 to 301 (1978 and 1979) of *The New England Journal of Medicine*[2-14] as examples. We highlight the clinically relevant issues for researchers and readers and refer them to the statistical literature[15-22] for mathematical details.

Parallel versus Crossover Design

In a two-treatment crossover study, each patient's response under treatment A is compared with the same patient's response under treatment B, so that the influence of patient characteristics that determine the general level of response can be "subtracted out" of the treatment comparison. This procedure does not remove biologic variation within an individual or measurement variation, but if these variations are small relative to the influence of patient characteristics, a crossover design based on a small sample of patients can provide the same statistical accuracy as a larger parallel study. Nevertheless, the decision to use a crossover design cannot be based solely on this potential saving in sample size, because powerful designs are also potential disasters. Their success balances on a narrow base of scientific and statistical assumptions. An investigator choosing between parallel and crossover designs should consider five factors that determine the effectiveness of the crossover design: (1) carry-over and period effects on treatment outcomes; (2) treatment sequencing and patient assignment; (3) crossover rules and timing of measurements; (4) dropouts, faulty data, and other data problems; and (5) statistical analysis and sample size.

Table 1 provides information relating to these factors for the 13 crossover studies we analyzed in the *Journal*. The studies were identified by means of a classification system developed for clinical studies. We classified a study as a crossover study only if both treatments were realistic candidates for clinical use and if each could be administered after the other. Therefore, a self-controlled study of a single treatment, in which treatment and control readings were taken, could not have been classified as a crossover study unless the use of no treatment at all was a realistic clinical alternative. Also, many medical/surgical studies do, in a sense, have crossovers from medical to surgical therapy, but these generally compare immediate surgery with an approach calling for medical treatment followed if necessary by surgical treatment. We would classify such studies as parallel comparisons. One variety of crossover design that we did not encounter in this series matches paired organs, such as teeth or limbs. The strengths and weaknesses of the paired-organs design are similar to those of an ordinary crossover design.

Table 1 indicates that the crossover design was used to study a wide variety of diseases and treatments,

From the Department of Biostatistics, Harvard School of Public Health, and the Department of Psychiatry, Massachusetts General Hospital. Address reprint requests to Dr. Louis at the Department of Biostatistics, Harvard School of Public Health, 677 Huntington Ave., Boston, MA 02115.

Supported by a grant (RF 79026) from the Rockefeller Foundation.

The New England Journal of Medicine is produced by NEJM Group, a division of the Massachusetts Medical Society.
Downloaded from nejm.org at Louisiana State Univ Health Sciences Ctr on August 28, 2025.
Copyright © 1984 Massachusetts Medical Society. All rights reserved, including those for text and data mining, AI training, and similar technologies.

Table 1. Characteristics of 13 Crossover Studies in the *Journal*, 1978-1979.*

Article Reference No.	No. of Patients	Method of Allocation	Crossover Rule	Blind Crossover	Assessed Order Effects	Used Multivariate Methods	Disease and/or Patients	Treatments	Outcome Variables
5	11	Rand	Time	No	No	Yes	Insulin-dependent diabetics, 20–29 years old	Subcutaneous insulin injection, into arm, leg, or abdomen, each with rest or exercise	Insulin absorption; plasma glucose levels
7	9	Rand	Time	Yes	No	No	Healthy non-coffee drinkers	(a) Placebo (b) Caffeine	Plasma renin; catecholamine; cardiovascular control
12	17	Rand	Time	No	No	Yes	Moderate to severe asthma	Seven different combinations of terbutaline, aminophylline, and placebo	Spirometric measurements and cardiovascular function
9	10, 21	Rand	Time	No	No	Yes	Irritable-bowel syndrome	Clidinium, placebo (normal controls untreated)	Myoelectrical and motor activity of colon
2	10	Rand	Time	No	No	Yes	Classic stable exertional angina	Passive smoking with or without ventilation; nonsmoking control	Exercise tolerance
10	12	Rand	Time	Yes	No	Yes	Men with stable angina pectoris	Five beta-adrenoreceptor–blocking agents, placebo; each at 4 doses	Cardiac function
4	113	Rand	State	No	No	Yes	Patients receiving cancer chemotherapy	(a) Prochlorperazine (b) Nabilone	Nausea; vomiting; side effects
13	22	Alt	Time	No	Yes	Yes	Severe burns or trauma	Three formulations for intravenous nutrition	Degree of protein catabolism
6	4	SFS	Time	No	No	Yes	Juvenile diabetes	Two diets, each with three drug treatments	Several metabolic measurements
11	20	SFS	Imp	No	No	Yes	Adults with asthma	Theophylline in various formulations	Serum concentration of theophylline
3	6	SFS	Time	No	No	Yes	Thalassemia	Drinking tea vs. water	Iron absorption
8	5	SFS	Time	Yes	No	Yes	Systemic mastocytosis	(a) Disodium cromoglycate (b) Lactose	Histamine in urine; symptoms recorded in diaries
14	15	Hap	Imp	No	No	No	Children with chronic asthma	(a) Alternate-day prednisone (b) Beclomethasone propionate (c) Combination	Hypothalamic–pituitary–adrenal function

*Rand denotes randomized, Alt alternating, SFS same fixed sequence, and Hap haphazard. Time denotes time-dependent, State state-dependent, and Imp impossible to determine.

with responses generally measured by a biochemical marker. The diseases were stable, and the treatment effects transient, but this group of studies generally did not satisfy basic design requirements. For example, only 7 of the 13 employed random allocation of treatment order, and only one included even a basic assessment of the effect of treatment order on outcome. Although 11 of the reports included acceptable statistical analyses, each failed to report some potentially important information.

Rules calling for time-dependent, as opposed to disease-state–dependent, crossover are necessary for a valid treatment comparison and were used in 10 of the 13 studies. Only 3 of the 13 studies blinded the crossover point, but blinding was apparently impossible in 4 of the remaining 10.

The conclusions drawn in these studies could have been strengthened by empirical evidence that lasting effects of one treatment were unlikely to influence the results of subsequent treatments. The assumption that there were no residual effects can be supported only by biologic principles or previous laboratory and clinical data. The degree of credence appropriate to the reported results depends heavily on the relevance and persuasiveness of such information.

Example

The following example illustrates the power of the crossover design and the potential problems that are associated with it.

With just four diabetic patients, Raskin and Unger[6] were able to compare the effects of three insulin-infu-

The New England Journal of Medicine is produced by NEJM Group, a division of the Massachusetts Medical Society.
Downloaded from nejm.org at Louisiana State Univ Health Sciences Ctr on August 28, 2025.
Copyright © 1984 Massachusetts Medical Society. All rights reserved, including those for text and data mining, AI training, and similar technologies.

sion regimens on chemical components of blood and urine. As part of the experiment, they monitored urea nitrogen excretion for 48 hours — first while patients were receiving intravenous insulin and somatostatin and again after switching the same patients to intravenous insulin, somatostatin, and glucagon. Our Table 2 is adapted from their Table 1. The data shown there represent the rate of urinary excretion of urea nitrogen in each of four patients while they were receiving insulin and somatostatin and again when they were receiving insulin, somatostatin, and glucagon.

If these eight measurements had been obtained in eight patients in a study with a parallel design, the difference in mean nitrogen excretion rate (3 g per 24 hours) would not have been statistically significant. The standard error of the difference in means would be 2.76, computed from the insulin and somatostatin and the insulin, somatostatin, and glucagon columns in an unpaired manner (see Table 2). Using instead the within-patient differences as the basis of the data, we obtain exactly the same mean difference (3 g per 24 hours) but with an S.E. of 0.4. These differences produce a paired t-statistic of 7.5 with 3 degrees of freedom, providing strong evidence for the hypothesis that the change from insulin and somatostatin in the first period of the experiment to insulin, somatostatin, and glucagon in the second period raised the level of nitrogen excretion.

These results could be interpreted either as supporting a difference between the effects of insulin and somatostatin on the one hand, and insulin, somatostatin, and glucagon on the other, or as supporting a difference produced by the order of administration (the combination of insulin and somatostatin was always used first). The study provides no information on or control for the effects on response of treatment sequence or of changes in disease state. Although diabetes is a fairly stable disease and the investigators incorporated a washout period, a stronger study would have resulted if they had treated only half the patients first with insulin and somatostatin and the other half first with insulin, somatostatin, and glucagon. Patients could have been selected on the basis of low urea nitrogen values, and regression to the mean (the tendency of unusually high or low values to become less extreme) could have accounted for the observed difference. Also, no matter how powerful the design, studies based on data from only four patients may have relatively little clinical impact until they are confirmed with larger samples.

If the order of treatments made no difference and the variation in response remained as in Table 2, a parallel comparison would require about 14 times as many patients to achieve the same level of statistical significance (about 56 patients divided equally between treatment groups). Even if no reduction of variance were obtained, the crossover design would require only half the number of patients to produce the same precision as a parallel comparison. Each patient would contribute information on both treatments, but each would be under study longer than in a parallel design.

Table 2. Urinary Excretion of Urea Nitrogen in Four Diabetic Patients.*

Patient No.	Treatment		Difference †
	IS	ISG	
	g of urea nitrogen/24 hr		
1	14	17	3
2	6	8	2
3	7	11	4
4	6	9	3
Mean	8.25	11.25	3.00
S.E.M.	1.90	2.00	0.40

*Data are adapted from Raskin and Unger.[6] IS denotes intravenous insulin and somatostatin, and ISG intravenous insulin, somatostatin, and glucagon.

†The S.E. of the difference between the means of ISG and IS: $2.76 = \sqrt{(1.90)^2 + (2.00)^2}$, if the groups were unpaired.

This reduction in sample size can be crucial in the study of treatment for an uncommon condition with a long and complex course. For example, Soter et al.[8] evaluated the use of disodium cromoglycate in the treatment of five patients with systemic mastocytosis. Graphs show the time course of symptoms in relation to the initiation and withdrawal of several administrations of drug and placebo. The high degree of patient-to-patient variability and the small number of available patients necessitated the use of the repeated crossover design to provide credible information on treatment efficacy.

Key Factors

The five factors listed earlier as determining the effectiveness of the crossover design need to be considered carefully before any crossover study is mounted. We can give some general guidelines on their consideration, and in this section we continue our discussion of them as they apply to the 13 studies listed in Table 1.

Carry-over and Period Effects

Carry-over Effects

The therapeutic effects of the first treatment may persist during the administration of the second. Investigators can often minimize this influence by appropriately delaying treatment administration. For example, Thadani et al.[10] compared five beta-adrenoreceptor blockers for their immediate effects on pain and other clinical signs in patients with stable angina. They administered and evaluated the drugs at weekly intervals, allowing time for the effects of one drug to dissipate before administering the other.

Period Effects

The disease may progress, regress, or fluctuate in severity during the period of investigation. For example, in the study of five patients with systemic mastocytosis, Soter et al.[8] dealt with the complex course of the disorder by switching treatment several times for each patient, with different intervals between the

The New England Journal of Medicine is produced by NEJM Group, a division of the Massachusetts Medical Society.
Downloaded from nejm.org at Louisiana State Univ Health Sciences Ctr on August 28, 2025.
Copyright © 1984 Massachusetts Medical Society. All rights reserved, including those for text and data mining, AI training, and similar technologies.

crossover points. The effects of any systematic trends or cycles should have balanced out in the design.

These kinds of influences are called order effects. They complicate the task of interpretation and analysis, and weaken the scientific and statistical basis for choosing a crossover design. Carry-over effects are the most troublesome, for they suggest that a subsequent treatment's activity depends on the previous treatment. With proper design, both types of order effects can be assessed and removed from the treatment comparison. The assessment of order effects must be based on a statistical model. If an order effect is comparable to or greater than the treatment effect, the different treatment sequences should be considered distinct test regimens. In this situation, the design is no more powerful than that of a standard parallel comparison and can be less powerful or invalid. Brown[15] shows that estimating and adjusting for order effects requires a sample size greater than that needed for a parallel comparison. Therefore, unless order effects are known to be negligible, the crossover design loses its advantages.

Order effects may have operated in each of the 13 crossover studies in our sample. As shown in Columns 3 through 5 of Table 1, investigators tried to various extents to control order effects by balanced sequencing and carefully timed measurements, but only 1 of the 13 reported on these attempts (Column headed "Assessed Order Effects"). The reasons for this low reporting rate are not clear. Lack of carry-over effects seems probable in most of the studies, but period effects could have been present. Readers of these reports could evaluate them more easily if the authors had included additional detail, such as an estimate and confidence interval for the treatment according to period interaction.[18,21] This interaction summarizes all order effects that may be present. A small interaction relative to the treatment effect validates the qualitative conclusions of the study. A descriptive summary of the interaction can be provided by a simple 2-by-2 table cross-classifying average treatment response and period; graphic displays are discussed in Huitson et al.[22]

Crossover designs are most appropriate in the study of treatment for a stable disease. Stable situations are studied in 12 of the 13 *Journal* articles we examined. In the one exception[13] the authors took special care to prevent bias resulting from the rapid changes in patient status.

Treatment Sequencing and Patient Assignment

The investigator must assign each patient to an initial treatment, and if there are more than two treatments or more than one administration of the treatments, he or she must specify the sequence. If all patients receive treatment according to the same fixed sequence, A followed by B, comparisons must be based on the assumption that the effects of the second treatment (B) after the first (A) do not differ from the effects B would have if it were given first. Such an experiment would provide no data by which to assess this assumption. If disease or treatment characteristics make B after A a fundamentally different treatment from B alone (e.g., chemotherapy after radiation vs. chemotherapy alone), no treatment comparison can be made. If some patients receive the sequence AB and others BA, then information is available on this issue.

If patients present for study over time, there are four basic ways of assigning them to treatment sequences: (1) by use of the same fixed sequence for all patients (four papers in our sample); (2) by random assignment among the sequences (seven papers); (3) by deterministically balanced assignment — for example, giving the first patient AB and the second BA, then repeating as often as needed (one paper); and (4) by uncontrolled, haphazard assignment — using sequences neither fixed in advance nor governed by a randomization procedure (one paper).

Method 1 does not provide the information needed to estimate and adjust for order effects. With this method, the validity of treatment comparisons depends on the near absence of order effects, and such absence would have to be established by data or argument external to the experiment.

Random assignment protects against conscious and unconscious bias. Deterministic balancing and haphazard assignment may be as valid but are more prone to biases caused by selection of study participants and initial treatment assignments. Therefore, we recommend random assignment of treatment order, with forced balancing of groups of patients when necessary (blocking; see Lavori et al.[1]).

Crossover Rules and Timing of Measurements

A previously specified crossover rule strengthens the scientific and clinical validity of a study. Investigators commonly employ either of two types of crossover rules — one that calls for a switch in treatments after a specified length of time ("time dependent") and one in which a crossover occurs when indicated by the clinical characteristics of the patient ("disease-state dependent"). These rules have different impacts on the magnitude and interpretation of order effects and on the general scientific strength of the study.

As with other aspects of designed experiments, the timing of measurements should be explicitly incorporated into the research protocol. The most scientifically acceptable switch points depend only on elapsed time. Initiating a treatment in response to the appearance of symptoms and withdrawing it or changing treatments when symptoms disappear makes it difficult or impossible to interpret observed treatment effects. In the 13 papers under study, 10 had time-dependent crossover rules and 1 a disease-state–dependent rule; in 2, the rules governing crossover were not clearly reported.

Whenever possible, crossover points should be concealed from patients and observers (blinded). Knowledge of a switch can influence treatment response or

The New England Journal of Medicine is produced by NEJM Group, a division of the Massachusetts Medical Society.
Downloaded from nejm.org at Louisiana State Univ Health Sciences Ctr on August 28, 2025.
Copyright © 1984 Massachusetts Medical Society. All rights reserved, including those for text and data mining, AI training, and similar technologies.

assessment or both, so that blinding the crossover point can reduce the influence of order effects.

Dropouts — Faulty and Outlying Data

Although dropouts and implausible data points are problems for any study, their effects may be pronounced in a study with a crossover design, because each patient contributes a large proportion of the total information and the design is sensitive to departures from the ideal plan. For example, consider again the study of diabetic patients[6] and the data in Table 2. If Patient 1 drops out during the first treatment period, only three patients are left to provide useful data. The mean difference between the two treatment regimens is still 3.00 g, but its standard error is now 0.577, and the t-statistic with 2 degrees of freedom equals 5.2, with a P value of 0.03. Although the result is still statistically significant, the P value has risen dramatically from the original 0.006. If the initial result had been less definitive, this loss of one patient would have altered the conclusions of the study. A single dropout in the comparable parallel study (with 28 patients in each group) should have had little influence on the conclusions.

Dropout rates can be high in crossover studies, since patients must receive at least two treatments to provide a complete data point. Partial information on a patient completing one treatment and then dropping out can be used in estimating the treatment effects in Period 1, provided that dropping out is not related to treatment response. A high dropout rate greatly weakens the study, and the initial sample size should be sufficiently large to compensate for this effect. Only 1 of our 13 studies reported having a subject drop out after the initial assignment of therapy. These studies were generally short and used treatments with few extreme side effects — the ideal combination for a crossover design.

Statistical analyses should include the identification of dropouts and deviant data points, and conclusions should not be sensitive to the latter. Protection against such sensitivity can be obtained either by setting aside deviant data or by using new techniques developed to be "resistant."[23,24]

Statistical Analysis and Sample Size

Observations made repeatedly in the same patient tend to be more similar than those made in different patients. Statistical analyses that take this relation into account are more complicated but potentially more powerful than those that are appropriate for a parallel comparison. Most important, the patient and not an individual measurement is the basic unit of statistical analysis.

Proper statistical analysis begins in effect by comparing data from a single patient over time and then combines these comparisons across patients. When only two measurements are compared (as in the Raskin and Unger study[6]), the paired t-test provides a widely applicable procedure. In many studies, however, three or more observations are generated by each patient, and a different statistical analysis is necessary.

For example, in the passive-smoking study,[2] each patient generated six measures of the length of time to angina (a base-line and post-treatment measurement for each of three treatment days). These data were collapsed to form three values by computing the difference between post-treatment and base-line measurements for each day. A fully multivariate analysis would have used these three values as a single (though still multivariate) unit for analysis. Analysis should proceed with the use of the multivariate regression and analysis of variance techniques discussed in his Chapter 5 by Morrison,[25] and by Grizzle,[17] Hills and Armitage,[18] and Layard.[20] These techniques model the association (correlation) among the measurements for a single individual and use this association in computing standard errors of comparisons. Also, the techniques allow adjustment of P values for multiple tests on the series of measurements. In effect, they operate by linking together the results of several paired t-tests.

In our sample of 13 studies, 2 used no multivariate methods, and 11 employed them with various degrees of sophistication to assess pair-wise treatment comparisons. All authors failed to exploit the data structure fully. Billewicz[26] reported a similar failure in 9 of 20 examples from the medical literature.

Of the 13 crossover studies, 12 used a small number of patients (4 to 22), though the total number of observations of treatments was much larger. This economical use of patients both justifies and results from the powerful crossover design.

Further Comparisons with Parallel Designs

If the most favorable underlying conditions for a crossover study are met (that is, if neither time nor treatment order affect the response to the currently administered treatment), pairing each patient with himself or herself and analyzing the data properly eliminate the influence of patient characteristics on the treatment comparison. This pairing represents the ultimate form of statistical adjustment for such characteristics.[1]

Even when the disease and treatments are satisfactory for a crossover trial, the choice between such a trial and a parallel comparison can be difficult. Although a crossover design has the potential advantage of economy and of providing a direct comparison of treatments in the same patient, the parallel design allows more straightforward analysis, its efficacy is less dependent on assumptions about the disease process, and it generally produces a lower dropout rate because each patient generates fewer measurements in a shorter time. In addition, the use of base-line measurements and statistical adjustments can greatly increase the precision of a parallel design.

For an example of the potential increase in precision

The New England Journal of Medicine is produced by NEJM Group, a division of the Massachusetts Medical Society.
Downloaded from nejm.org at Louisiana State Univ Health Sciences Ctr on August 28, 2025.
Copyright © 1984 Massachusetts Medical Society. All rights reserved, including those for text and data mining, AI training, and similar technologies.

in a parallel design, let us look again at the insulin study,[6] but suppose that the data came from a study with a parallel design using four men and four women, as shown in Table 3. Notice that the levels of urea nitrogen excretion are 5.0 g higher per 24 hours in men than in women, but that the difference between the means for the two treatment regimens is still 3.0 g per hour. Although we do not have paired responses for four patients, we can use sex to explain some of the patient-to-patient differences in response. This explanation is accomplished by the analysis of variance, in which a level of urea nitrogen excretion is explained by the sum of a base-line value, an effect of treatment (insulin and somatostatin vs. insulin, somatostatin, and glucagon), an effect of sex (male vs. female), and a residual (the deviation between the observed response and that predicted by the model). See Table 3 for two examples.

Table 3. Variation on the Same Study Shown in Table 2, with a Hypothetical Sex Variable.*

	TREATMENT		MEAN	
	IS	ISG		SEX EFFECT
		g of urea nitrogen/24 hr		
Men	14 7	11 17	12.25	2.5
Women	6 6	9 8	7.25	−2.5
Mean	8.25	11.25	9.75	
Effect	−1.5	1.5		

*Response = base-line value + (treatment effect) + (sex effect) + residual; residuals (not shown above) are the deviations between observed responses and those predicted by the model. They make the right-hand side equal the left-hand side in the general formula and numerical examples.

Examples:	Base Line		Treatment		Sex		Residual
14 =	9.75	−	1.5	+	2.5	+	3.25
8 =	9.75	+	1.5	−	2.5	−	0.75

The standard error computed from this analysis of variance is 2.09 — a 24 per cent reduction from the 2.76 computed in Table 2. The t-statistic is 1.44 (3.00/2.09) with 5 degrees of freedom ($P = 0.20$). Recall that the unpaired comparison of treatment means produces a t-statistic of 1.08 (3.00/2.76) with 6 degrees of freedom ($P = 0.32$). Although the sex-adjusted analysis is still not statistically significant, paying this one degree of freedom to remove the sex effect makes the experiment more precise.

The use of sex to help explain the data is an example of covariance adjustment. A parallel-design study with a covariance adjustment for sex would have to be only about 11 times as large as the crossover study, instead of the 14-fold size mentioned above. Such covariance adjustments must be interpreted with care, especially if they are based on patient characteristics discovered through an exhaustive search to produce a desired result.

In a choice between a parallel and a crossover design, we think the burden of proof should be placed on those favoring the crossover design to show that it can succeed in improving on the parallel design, though we think that such proof will often be forthcoming. Evidence of a strong likelihood of success with a crossover design in a particular study would include previous studies validating the absence of carry-over effects, low dropout rates, and a relatively stable disease process. Even in such situations a parallel design adjusted for previously specified covariates can be a strong competitor. We recommend that a design be chosen only after careful consideration of research goals, the disease process, and the trade-off between power and fragility. Such consideration requires a collaborative effort among clinical, laboratory, and statistical scientists.

Patients as Their Own Controls

Volumes 298 through 301 of the *Journal* contained an additional 28 reports of studies in which patients served as their own controls before, during, or after treatment. Such research designs incorporate many of the features of crossover studies, but new problems arise, as we will discuss after we give three examples of self-controlled studies. In one example, Peck et al.[27] evaluated the effect of 13-*cis*-retinoic acid on severe acne by examining changes in disease status; it appears that the acne was of such duration and intractability in each patient that substantial spontaneous improvement was unlikely to interfere with treatment evaluation. In another example,[28] patients with hypertension that had been found on screening were followed to determine whether informing them that they had elevated blood pressure was reflected in increased absenteeism for illness. Each patient's work record after diagnosis was compared with his previous record. (Other parts of this study used crossover and parallel designs; we will not discuss those aspects here.)

Packer et al.[29] used a self-controlled design to investigate the clinical reaction of patients with congestive heart failure to the abrupt withdrawal of nitroprusside. Previous research had established that nitroprusside produces rapid hemodynamic and clinical improvement in patients with congestive failure, but some results had suggested that the sudden termination of treatment caused adverse clinical reactions. The self-controlled design seems ideal for the study of this problem, and it is hard for us to see how either crossover or parallel designs, usually considered more powerful, could have been employed.

Many crossover and parallel[1] studies include an element of self-controlling. We have classified such studies as crossover or parallel, and concentrate here on designs in which observations before, during, or after treatment are used as the main control on experimental variation.

Each of the five critical issues in crossover studies discussed above applies with undiminished force to self-controlled studies, and several new ones arise. We will discuss the following, though at less length than

The New England Journal of Medicine is produced by NEJM Group, a division of the Massachusetts Medical Society.
Downloaded from nejm.org at Louisiana State Univ Health Sciences Ctr on August 28, 2025.
Copyright © 1984 Massachusetts Medical Society. All rights reserved, including those for text and data mining, AI training, and similar technologies.

the first five: lack of symmetry; the nature of the problems studied; the study of patients with refractory disease; and the absence of direct comparisons.

Lack of Symmetry

A critical difference between crossover and self-controlled studies is the fundamental lack of symmetry in the latter between observations during treatment and those during control periods. These observations often differ in such matters as duration, nature, and intensity of clinical studies, and decision rules about when to modify or switch treatments. Such differences can lead to substantial bias (for example, in opportunities to observe untoward developments in the disease process) that may tilt results toward or away from the treatment under study. In well-designed crossover studies, this problem does not arise, because such effects are balanced between treatments and become part of the random error.

Nature of the Problems Studied

Another general difference between self-controlled and crossover studies is in the nature of the problems studied. In our sample, self-controlled studies were more often used at early points in the clinical development of new treatments, so that attention focused on multiple laboratory measurements rather than on one or two measures of clinical outcome. For example, Stacpoole et al.[30] used a self-controlled design in the first study in human subjects of the metabolic effects of dichloroacetate, which had been found to be effective in diabetic and starved rats. They measured blood glucose, plasma lactate, alanine, and other biochemical indexes before and after treatment. Similarly, Davis et al.[31] used this design to study the effects of captopril on arterial blood pressure, cardiac output, and plasma renin activity, and Gavras et al.[32] used it to investigate the effects of a new oral inhibitor of angiotensin-converting enzyme, SQ 14225, on cardiovascular status, liver function, and kidney function.

Although this measuring of multiple outcomes is not necessarily a weakness in self-controlled designs, it does mean that methods of analysis must be carefully tailored to exploit the data. One difficulty is the "multiple-comparisons problem," addressed by Ingelfinger et al.[33] and others. Another is the optimal multivariate analysis of numerous dependent, as well as independent, variables. Expert help in multivariate statistical methods was rarely evident in the 28 self-controlled studies that we reviewed, though we believe it could often have led to substantially more productive use of the data.

Study of Patients with Refractory Disease

A third common difference from crossover studies was the focus of many self-controlled studies (at least 7 of the 28 in our sample) on patients whose disease had responded inadequately to standard therapies. This may result, in part, from the use of this design at early stages of investigation in human subjects and the ethical considerations that lead to the enrollment of patients with refractory disease. We cite three examples. In a study by Bilezikian et al.[34] of mithramycin-induced hypocalcemia in patients with Paget's disease, the patients "had previously been treated with a variety of agents, including oral phosphate, calcitonin and ethane-1,hydroxy-1,1-diphosphonate, to which they had, in general, responded poorly." In the study of 13-*cis*-retinoic acid for acne,[28] patients "uniformly had a history of minimal response to treatment with oral and topical antibiotics, oral vitamin A, topical vitamin A acid, topical benzoyl peroxide, x-rays and other acne therapies." And a study by Tamborlane et al.[35] of portable infusion pumps for the injection of insulin focused on juvenile diabetic patients who had continued to excrete albumin while receiving intravenous insulin.

Several problems arise in studies of patients with refractory disease. One is that intense, prolonged (though perhaps incompletely effective) treatment may require a prolonged "washout" or recovery period, while the patient's clinical problems demand prompt relief. We found no evidence that clinical imperatives resulted in inadequate washouts in our sample, but it is not clear that any of the investigators looked for such evidence. Another problem results from "regression toward the mean" — the tendency of an extreme value when it is remeasured to be closer to the mean, because the original value was likely to have been unduly influenced by random variation. If patients are enrolled in a self-controlled study of some new regimen when standard treatments appear to be losing efficacy (that is, when the patients' conditions are worse than average), some general improvement may occur that has nothing to do with improved therapy. This problem can sometimes be addressed by using a first set of measurements after a washout period to establish patient eligibility for a self-controlled study and a second set, after a further "stabilizing" period, to establish base-line levels.

Patients whose disease fails to respond to standard treatment may include those with variant forms of the disease or those whose general conditions have deteriorated so that even a new, effective treatment is of little avail. Thus, such patients may not provide a fair test of the capabilities of a new treatment, but the value of studying such patients in the early development of a new treatment is often compelling. The main advantage is the reduced risk of compromising the well-being of patients by withholding standard treatments, which in these patients are already known to be of little effect. Although studies of nonresponders can rarely provide definitive conclusions about the efficacy of a treatment in more responsive disease, they can provide information of considerable value in planning further studies (e.g., randomized controlled trials) that are designed to give unambiguous answers to questions about broad treatment recommendations.

The New England Journal of Medicine is produced by NEJM Group, a division of the Massachusetts Medical Society.
Downloaded from nejm.org at Louisiana State Univ Health Sciences Ctr on August 28, 2025.
Copyright © 1984 Massachusetts Medical Society. All rights reserved, including those for text and data mining, AI training, and similar technologies.

Absence of Direct Comparisons

Self-controlled studies do not provide direct information about how a new treatment compares with standard therapies. Rather, one must combine the results of self-controlled studies with those of other studies, though this is often difficult because of inevitable differences in study design and conduct. The demographic composition of the patient groups may differ, or the disease may be more severe in one group than another. Often, one cannot even determine the direction of such differences because investigators have used subjective or partly subjective assessments of clinical status, or they have not used the same set of objective measurements to establish a diagnosis or to monitor disease progress. Different investigators may even use different outcome variables to assess the effects of treatment. One must adjust for all these matters to obtain a valid assessment of the relative merits of two treatments, though it is generally not clear how to do so, and there is rarely any way to tell whether the adjustments were appropriate and adequate. Thus, despite the clear value of self-controlled studies in the initial investigation of new treatments, one must usually use a randomized controlled trial or some other powerful research design to determine with assurance whether the new treatment should be recommended for general use.

We are indebted to the members of the *New England Journal of Medicine* Study Group for their comments and to Mary Schaefer for assistance in the preparation of the manuscript.

References

1. Lavori P, Louis TA, Bailar JC, Polansky M. Designs for experiments — parallel comparisons of treatment. N Engl J Med 1983; 309:1291-9.
2. Aronow WS. Effect of passive smoking on angina pectoris. N Engl J Med 1978; 299:21-4.
3. de Alarcon PA, Donovan M-E, Forbes GB, Landaw SA, Stockman JA III. Iron absorption of the thalassemia syndromes and its inhibition by tea. N Engl J Med 1979; 300:5-8.
4. Herman TS, Einhorn LH, Jones SE, et al. Superiority of nabilone over prochlorperazine as an antiemetic in patients receiving cancer chemotherapy. N Engl J Med 1979; 300:1295-7.
5. Koivisto VA, Felig P. Effects of leg exercise on insulin absorption in diabetic patients. N Engl J Med 1978; 298:79-83.
6. Raskin P, Unger RH. Hyperglucagonemia and its suppression: importance in the metabolic control of diabetes. N Engl J Med 1978; 299:433-6.
7. Robertson D, Frölich JC, Carr RK, et al. Effects of caffeine on plasma renin activity, catecholamines and blood pressure. N Engl J Med 1978; 298:181-6.
8. Soter NA, Austen KF, Wasserman SI. Oral disodium cromoglycate in the treatment of systemic mastocytosis. N Engl J Med 1979; 301:465-9.
9. Sullivan MA, Cohen S, Snape WJ Jr. Colonic myoelectrical activity in irritable-bowel syndrome: effect of eating and anticholinergics. N Engl J Med 1978; 298:878-83.
10. Thadani U, Davidson L, Singleton W, Taylor SH. Comparison of the immediate effects of five β-adrenoreceptor-blocking drugs with different ancillary properties in angina pectoris. N Engl J Med 1979; 300:750-5.
11. Weinberger M, Hendeles L, Bighley L. The relation of product formulation to absorption of oral theophylline. N Engl J Med 1978; 299:852-7.
12. Wolfe JD, Tashkin DP, Calvarese B, Simmons M. Bronchodilator effects of terbutaline and aminophylline alone and in combination in asthmatic patients. N Engl J Med 1978; 298:363-7.
13. Woolfson AMJ, Heatley RV, Allison SP. Insulin to inhibit protein catabolism after injury. N Engl J Med 1979; 300:14-7.
14. Wyatt R, Waschek J, Weinberger M, Sherman B. Effects of inhaled beclomethethasone dipropionate and alternate-day prednisone on pituitary-adrenal function in children with chronic asthma. N Engl J Med 1978; 299:1387-92.
15. Brown BW Jr. The crossover experiment for clinical trials. Biometrics 1980; 36:69-79.
16. Cole JWL, Grizzle JE. Applications of multivariate analysis of variance to repeated measurements experiments. Biometrics 1966; 22:811-28.
17. Grizzle JE. The two-period change-over design and its use in clinical trials. Biometrics 1965; 21:467-80.
18. Hills M, Armitage P. The two-period cross-over clinical trial. Br J Clin Pharmacol 1979; 8:7-20.
19. Kershner RP, Federer WT. Two-treatment crossover designs for estimating a variety of effects. J Am Stat Assoc 1981; 76:612-9.
20. Layard MWJ, Arvesen JN. Analysis of Poisson data in crossover experimental designs. Biometrics 1978; 34:421-8.
21. Wallenstein S, Fisher AC. The analysis of the two-period repeated measurements crossover design with applications to clinical trials. Biometrics 1977; 33:261-9.
22. Huitson A, Poloniecki J, Hews R, Barker N. A review of cross-over trials. Statistician 1982; 31:71-80.
23. Velleman PF, Hoaglin DC. Applications, basics, and computing of exploratory data analysis. Boston: Duxbury Press, 1981.
24. Mosteller F, Tukey JW. Data analysis and regression: a second course in statistics. Reading, Mass.: Addison-Wesley, 1977.
25. Morrison DF. Multivariate statistical methods. 2nd ed. New York: McGraw Hill, 1976.
26. Billewicz WZ. The efficiency of matched samples: an empirical investigation. Biometrics 1965; 21:623-43.
27. Peck GL, Olsen TG, Yoder FW, et al. Prolonged remissions of cystic and conglobate acne with 13-*cis*-retinoic acid. N Engl J Med 1979; 300:329-33.
28. Haynes RB, Sackett DL, Taylor DW, Gibson ES, Johnson AL. Absenteeism from work after detection and labeling of hypertensive patients. N Engl J Med 1978; 299:741-4.
29. Packer M, Meller J, Medina N, Gorlin R, Herman MV. Rebound hemodynamic events after the abrupt withdrawal of nitroprusside in patients with severe chronic heart failure. N Engl J Med 1979; 301:1193-7.
30. Stacpoole PW, Moore GW, Kornhauser DM. Metabolic effects of dichloroacetate in patients with diabetes mellitus and hyperlipoproteinemia. N Engl J Med 1978; 298:526-30.
31. Davis R, Ribner HS, Keung E, Sonnenblick EH, LeJemtel TH. Treatment of chronic congestive heart failure with captopril, an oral inhibitor of angiotensin-converting enzyme. N Engl J Med 1979; 301:117-21.
32. Gavras H, Brunner HR Turini GA, et al. Antihypertensive effect of the oral angiotensin converting-enzyme inhibitor SQ 14225 in man. N Engl J Med 1978; 298:991-5.
33. Ingelfinger J, Mosteller FM, Thibodeau L, Ware J. Biostatistics in clinical medicine. New York: Macmillan, 1983.
34. Bilezikian JP, Canfield RE, Jacobs TP, et al. Response of 1α, 25-dihydroxyvitamin D to hypocalcemia in human subjects. N Engl J Med 1978; 299:437-41.
35. Tamborlane WV, Sherwin RS, Genel M, Felig P. Reduction to normal of plasma glucose in juvenile diabetes by subcutaneous administration of insulin with a portable infusion pump. N Engl J Med 1979; 300:573-8.

The New England Journal of Medicine is produced by NEJM Group, a division of the Massachusetts Medical Society.
Downloaded from nejm.org at Louisiana State Univ Health Sciences Ctr on August 28, 2025.
Copyright © 1984 Massachusetts Medical Society. All rights reserved, including those for text and data mining, AI training, and similar technologies.