The randomized controlled trial is often difficult, impractical or unethical in the clinical setting. Specific types of experimental design are examined for application and ease of interpretation of results, with particular focus on the generalization and demonstration of cause and effect. Examples are given of relatively easy changes that greatly strengthen the design and several recently published studies are used as illustrative examples.

Your experimental design is a function of the question you ask, the goals you want to achieve with your research, the degree of scientific rigor you want to use to establish your results, and the resources you have available to conduct your study. Although the randomized, controlled trial is generally regarded as the strongest research design, this design is often difficult, impractical or unethical in the clinical setting. There are many types of designs available to the investigator as presented briefly in the first article in this series. The focus of this paper is on developing a stronger design in the clinical setting. This article will present several common designs that do not readily allow interpretation, and several simple modifications to these designs that greatly strengthen the conclusions that can be drawn. Finally, several reported studies will be examined for ways in which these aspects of design could have been incorporated without greatly increasing the complexity of the research.

Regardless of the area of study, the goals of research can be described as being either descriptive or inferential in nature. Descriptive research may seek, for example, to determine the average age of onset of multiple sclerosis (MS) by mailing a questionnaire to a large number of NIS patients and asking them to describe the age at which they first started to experience symptoms. Although information regarding average age of onset may be useful for further study, no conclusions other than the descriptive findings are possible from this type of research. Inferential research, on the other hand, seeks to draw conclusions from the data that are generalizable to other situations. Such conclusions may or may not include cause and effect.

An example of the distinction between causal and noncausal inferences might be illustrated by two different types of studies on the relationship between coffee drinking and stroke.

**Example 1**

The only conclusion that could be drawn from this study is that drinking more coffee is associated with a higher probability of having a stroke. Questionnaire data were gathered from a thousand individuals regarding their coffee-drinking habits. All sampled individuals were followed for 3 years and their coffee consumption and incidence of stroke were recorded. A statistically significant relationship between the amount of coffee and stroke was found.

The mere association between coffee drinking and stroke does not permit one to infer that coffee drinking is a causal factor in stroke. Other variables such as stress, hypertension, weight, other dietary factors, cigarette smoking, etc. might be the real causes and strongly related to coffee drinking. Unless these factors are taken into consideration in the study the results make it appear as though coffee is the factor responsible for stroke. Consider, however, the following example.

**Example 2**

One thousand subjects were randomly assigned to one of two conditions. In the first condition the 500 subjects were told to eliminate coffee from their diets and were followed for 3 years during which time they were monitored to determine the amount of coffee consumption and stroke. The 500 subjects in the second condition were given no special instructions but were followed for a period of 3 years and were also monitored for coffee consumption and stroke. At the end of the 3-year period, the results showed that the first group differed significantly from the control group, both in the average coffee consumed and in the number of strokes that occurred.

In this second study a randomized experiment was conducted; it is reasonable to infer that coffee drinking is a causal factor in stroke incidence because the only difference between the two groups was the amount of coffee consumed.
Differences on other variables have been controlled through the random assignment procedure. The number of subjects and follow-up effort is the same as in Example 1, but much more information can be drawn because of the addition of the random assignment into two groups and the elimination of coffee intake in one group.

This second type of research is referred to as a randomized experiment and is the best way to reach conclusions regarding causation. Although true random experiments are the best way to establish causation, there are many other ways to conduct experiments in applied settings that allow causal inferences to be made. Before discussing these methods, we should briefly review three typical research designs that do not allow causal inferences to be made. These designs include the one-group posttest, the posttest with nonequivalent groups and the one-group pretest-posttest.

**NONINTERPRETABLE DESIGNS**

**The One-Group Posttest**

One group posttest is a common design in research settings and is characterized by treatment followed by an observation or measurement.

*Example 3*

Fifteen low-level spinal cord-injured patients were given functional electrical stimulation (FES) for 0.5 h every day for 6 weeks. After treatment, patients were given standard sensory and motor tests. Patients reported that FES had a beneficial effect on their condition.

Although conclusions are often drawn from studies of this type, no real causal inference can be made. It is not clear whether patients improved because of FES or because of other rehabilitation treatments that were ongoing at the same time, or if, indeed, they actually improved at all. Such research is at the level of a case study and may be useful for generating hypotheses, piloting treatment programs or to obtain descriptive information regarding the situation in question. It does not provide direct information on whether the treatment actually caused the desired outcome.

**Posttest Design with Nonequivalent Groups**

Experimental Group | Treatment | Measure
--- | --- | ---
Control Group | Measure

This type of design is often utilized as an afterthought after a treatment program has been started.

*Example 4*

The effectiveness of a new occupational therapy treatment program was compared with the effectiveness of the previous program. Records were searched for patients who had been through the previous program to determine effectiveness of job placement and activities of daily living. Similar data were collected for patients who had been through the new program. Based on the number and quality of job placements, the new program was found to be significantly more effective. Although comparisons were made between two groups, the conclusions that were drawn might not be valid. There may be preexisting differences between the two groups that might have led to differences in job placement and daily living activities regardless of the occupational program. Without knowing whether the two groups differed on important details, it is difficult to draw clear conclusions. This design is commonly seen with the statement experimental and control groups did not differ significantly. " For example, if the mean age of patients in the old program is 60 and in the new program, 55, this difference may not be statistically significant with a small sample but the clinical effect on occupation is obvious.

**One Group Pretest-Posttest**

Measure 1 / Treatment / Measure 2

Another commonly used design involves the measurement of one group of subjects on some key variables before treatment and again after treatment.

*Example 5*
The research study described in Example I was conducted again with a new group. Sensory and motor tests were given before the FES program and again after the FES program. Differences between the pre-program results and post-program results are examined and the conclusion is reached that FES produced positive changes in muscle function for most of the subjects.

Although this study is somewhat more sophisticated than the design used in Example 1, it is still not possible to draw clear inferences that the results were actually the result of the treatment provided. The positive changes could have occurred naturally or might have been the result of some other aspects of therapy being conducted at the same time. Because there was not a comparable group that did not receive the treatment, conclusions regarding the effectiveness of FES cannot be drawn.

INTERPRETABLE QUASIEXPERIMENTAL DESIGNS

Although it is difficult to conduct strict randomized experiments in applied settings it is frequently possible to approach the level of control afforded by these ideal designs by using any of a number of quasiexperimental designs. These designs are referred to as quasiexperimental because they involve comparisons between groups that are not randomly determined but still allow reasonable causal interpretations to be made by controlling for other possible explanations of the results.¹

Although quasiexperimental designs vary considerably in complexity, four relatively simple designs are especially suited to research in applied settings. These include untreated control group design with pretest and posttest, removed treatment design with pretest and posttest, the repeated treatment design and cohort designs.

Untreated Control Group Design with Pretest and Posttest

<table>
<thead>
<tr>
<th>Experimental group</th>
<th>Measure 1</th>
<th>Treatment</th>
<th>Measure 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control group</td>
<td>Measure 1</td>
<td>Measure 2</td>
<td></td>
</tr>
</tbody>
</table>

Often, it is possible to find a naturally occurring comparison group to serve as a control for a treatment group.

Example 6

A study was conducted to examine the effects of a new drug on spasticity. Standard measures of spasticity were taken on 30 patients with spinal cord injury (SCI) 1 month post-injury after which treatment with the new drug was begun. After 1 month, spasticity was again measured on the 30 patients. Because standard measures of spasticity had been routinely taken on a monthly basis for the past 3 years, it was possible to identify, through archival records, 240 SCI patients not receiving treatment. Of the 240 in the records, a group of 30 was selected to match the treatment group in age, gender, level of injury and spasticity. When the treatment group was compared with this control, it was found that the level of spasticity was significantly lower 1 month after treatment with the new drug.

The study described in Example 6 illustrates how archival data can be used to find reasonable quasiexperimental controls. In this case, it was possible to "create" a control group matched on variables, especially the key variable of spasticity. Many opportunities exist for two group studies with pretest and posttest. Where possible, groups should be matched on the dependent (pretest-posttest) variable and any other variables known to be relevant to the outcome variable. Even when matching is not possible, however, the design can lead to useful conclusions because the posttest v pretest differences can be measured for both groups. A technique known as analysis of covariance can be used to statistically control for differences between groups.

Removed Treatment Design with Pretest and Posttest

<table>
<thead>
<tr>
<th>Measure 1 / Treatment / Measure 2 / Measure 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Remove Treatment / Measure 4</td>
</tr>
</tbody>
</table>

In cases where it is not possible to create a control group from archival data it may still be possible to infer causality by using one group as its own control. The removed treatment design involves a one-group pretest-posttest design, followed by a repeated pretest-posttest study on the same group during a time in which the treatment is removed.

Example 7
The effectiveness of the use of transcutaneous electrical nerve stimulation (TENS) in managing pain was investigated with a group of 25 patients. Subjective measures of pain were obtained from each patient after which a TENS protocol was followed for 2 weeks. At the end of the 2 weeks, the subjects were again measured for subjective pain. The TENS was continued for another week and a third measurement of pain was taken after which the TENS unit was altered so that a placebo treatment was given for 2 weeks. After this 2-week period, a final measurement of pain was taken. It was found that subjective reports of pain decreased significantly at the time of the second measurement and increased significantly at the time of the fourth measurement. It is concluded that the TENS treatment resulted in a significant lessening of pain for the subjects.

In this study, the causal inference that TENS reduced pain was reasonable because of the removed treatment condition. Because patients were unaware of the placebo treatment, it can be assumed that the increased pain was the result of the removal of the TENS stimulation. Without the removed treatment component of this study, we would have been left with a one-group pretest-posttest study; it would have been difficult to determine whether the results were the result of the TENS or other factors occurring during the 2-week study period.

Repeated Treatment Design
Measure 1 / Treatment / Measure 2 / Removal
Measure 3 / Treatment / Measure 4

The repeated treatment design takes the previous design one step further and is suitable when the effects of a treatment are unlikely to result in permanent change or improvement.

Example 8

As an alternative to the study described in Example 7, a single group of 25 subjects was measured for pain and then treated with TENS for 2 week after which subjective pain was measured a second time. After the second measurement, a placebo treatment was continued for 2 weeks after which a third pain measurement was taken. After the third pain measurement, all subjects were then given active TENS for 2 weeks. A fourth and final measurement was then taken. The results showed that there was a significant improvement between the first and second measurements, a significant increase in pain between the second and third measurements and a significant decrease in pain between the third and fourth measurements.

This alternative design allows even stronger conclusions to be drawn from a single research sample. Two notes should be added to the designs illustrated in Examples 7 and 8. Where possible, measurements should be taken at equal intervals and care should be taken that subjects and researchers are not influenced by knowledge of the nature of the treatment and the expected outcomes. Ideally, such research should be double-blind; both subject and researcher should be unaware of the true treatment v placebo condition.

Cohort Designs

<table>
<thead>
<tr>
<th>Cohort Group 1</th>
<th>M1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cohort Group 2</td>
<td>M2</td>
</tr>
<tr>
<td>Cohort Group 3</td>
<td>M3</td>
</tr>
<tr>
<td>Cohort Group 4</td>
<td>T</td>
</tr>
<tr>
<td>Cohort Group 5</td>
<td>T</td>
</tr>
<tr>
<td>Cohort Group 6</td>
<td>T</td>
</tr>
</tbody>
</table>

where M is measurement and T is treatment.

An especially powerful and potentially fruitful design in applied settings involves the study of cohort groups. In this context, cohorts can be defined as groups that enter and leave an institution (e.g., a rehabilitation center) at approximately the same time. Although many different types of cohort designs are possible a specific example may illustrate some of the virtues of this approach.

Example 9

Researchers at a rehabilitation center were interested in knowing the effectiveness of a training program designed to teach stroke patients and their families how to effectively manage their post-hospitalization care. The program was begun during the first quarter of 1986. Through its historical records, the hospital was able to identify several post-hos-
pitalization outcome measures for all patients beginning in 1984. These include reported stress in family members, independence in activities in daily living and recurrence of stroke. Cohort groups were identified by taking all stroke patients entering during each quarter from 1984 through 1986 making a total of 12 cohort groups. Records were obtained to determine the status of each patient on the outcome variables for 2 years after hospitalization; a summary index was used to summarize the status of each patient on all outcome variables. The graph shown in Figure 1 is the result. With the introduction of the program, there was a clear and consistent improvement in the summary index. However, there was also a gradual improvement over time even before the new program was instituted. Hence, all of the difference in outcome between 1984 and 1986 is not solely the result of the new program.

Cohort designs have several advantages. Although the groups are not randomly determined, they should be comparable, especially if several cohorts are studied under each condition. In Example 9, there were a total of eight "control" cohort groups and four "treatment" cohort groups. The clear and consistent difference between groups receiving the special program and groups not receiving the special program presents persuasive evidence for a causal interpretation provided other possible causal factors did not occur simultaneously. The fact that 12 groups were studied over a period of 3 years argues against the explanation of some alternative causal agent.

The designs presented here are only a few examples of how researchers can take advantage of available controls to reach more valid conclusions about new programs or forms of treatment. Specific illustration of this is given in three examples drawn from recent publications in the *American Journal of Physical Medicine and Rehabilitation*.

**Case 1**


In this study, 15 children who were using leather/metal braces were provided new plastic/metal braces and then new leather/metal braces. After at least 2 weeks with each brace, subjects were brought into the laboratory and gait analysis performed with each brace. Six months after the gait analysis, subjects were interviewed about their subsequent brace use and preference. As the order of assessment in the laboratory was randomly varied, this design can be schematically represented as:

<table>
<thead>
<tr>
<th>Acclimatization</th>
<th>Laboratory Testing</th>
<th>Follow-up</th>
</tr>
</thead>
<tbody>
<tr>
<td>P P L L M1 P M2</td>
<td>M3</td>
<td></td>
</tr>
<tr>
<td>P P L P M1 L M2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

where P is plastic, L is leather, M is measurement and 1 and 2 refer to use of plastic or leather braces.
The design could be strengthened by adding a measure of gait assessments in the old brace as a pretreatment measure. If this were taken at the same time as the other gait assessments, the complexity of the study would only be increased by the time required for a third gait assessment. In this case, the use of the old brace could be considered as a removal of treatments 1 and 2. Because the old braces were of variable quality, the results of preference of plastic over leather could be the result of the poor fit initially, not to any inherent difference between plastic and leather. However, without the initial measure, it is difficult to draw any conclusions as to why the children preferred one brace as "overall objective differences between orthosis type were minimal." A practical reason, which may have precluded this approach, is the fact that, for some of the children, almost 3 years elapsed between the original provision of the plastic brace and the gait assessment. By that time, the original brace probably did not fit. Therefore, some streamlining in the provision of the braces would also have been necessary to take this approach.

**Case 2**


The purpose of this study was to "... determine if geographic segregation of the treatment team within a specific area of the hospital would improve the team's performance in documenting patient's progress." A retrospective chart review was used to analyze 7 charts from the 2-month period immediately before the start of an arthritis treatment unit, 10 charts from 0-5 months after its onset and 10 charts from 18-24 months of operation. This can be schematically represented as:

Measure 1 / Treatment / Measure 2 / Treatment
Measure 3

Unfortunately, the study did not clearly define what the treatment was. Was it just geographic segregation of the patients or were there also changes in team meetings, goals and composition. Any program development such as this requires a great deal of preparation, so it is likely that there were some effects of this preparation even in the few months before it started. For example, if staff members had met to clarify their treatment goals, then they may have begun to state these goals in their care plans even before the program started. Likewise, it is not clear when the criteria for the chart audits were developed. Although staff members were not told the results of the audits, they were told what the evaluation criteria would be. Informing staff members of an impending audit and the criteria to be used is an excellent way to improve compliance. A better method would be to begin the audit procedure first, explaining to staff in detail the expectations for documentation before beginning the geographically isolated unit. After 6 months to 1 year, the unit could then be started to see its additional effect. This is, of course, difficult to do retrospectively, so an alternative would be to examine multiple cohorts before and after beginning the unit:

![Diagram](image)

where Treat is treatment and M is measurement.

**Case 3**


This study of 35 rehabilitation inpatients 2-10 months after traumatic brain injury was undertaken "... to define which subgroup of patients develop severe agitation warranting [drug] intervention and to ... predict therapeutic responsiveness to tricyclic antidepressants.” Fifteen subjects showed severe agitation not controlled by nonpharmacological methods and required the use of physical restraints. These patients were placed on one of two
medications (allocated randomly) with 10 receiving one drug and 5 the other. Ten of these 15 showed improvement, defined as a >50% decrease in number of agitated episodes.

This study illustrates several important principles. First, with small numbers of subjects, randomization between two groups is very likely to lead to imbalance (in this case 10 and 5) in the numbers of subjects in each group. There are techniques known as blocked randomization in which the numbers of subjects assigned to each group are equalized at certain intervals (e.g., after every 2, 4 or 10 subjects). This is done statistically before the project is begun, so that the investigator remains blind to assignment, but the maximum difference in group size is carefully controlled.

Second, the major outcome measure, number of agitated episodes per week, was changed from a continuous variable to a categorical one (±50% reduction from baseline) with consequent loss of information and power to detect smaller effects of the treatment.

Third, patients selected for the drug treatment were already extreme in the number of agitation episodes. In any study, patients at the extreme of measure can be expected to be less extreme on the next time period by a phenomenon called "regression to the mean." This results from the fact that a certain number of extreme values occurred just by chance and will not occur again. Thus, by selecting patients with the worst problem, a certain number will show scores which are improvements even though no basic change in patient status has occurred. This is further complicated by the natural tendency of persons with head injury to improve over time. Following the non-treated patients longer and discontinuing drug treatment would strengthen the study design although practical patient management considerations might make this difficult. The design is shown with suggested changes in bold:

<table>
<thead>
<tr>
<th></th>
<th>M1</th>
<th>M2</th>
<th>M3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-treated</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group A</td>
<td>M1</td>
<td>Drug A</td>
<td>M2</td>
</tr>
<tr>
<td>Group B</td>
<td>M1</td>
<td>Drug B</td>
<td>M2</td>
</tr>
</tbody>
</table>

EXERCISES

1. Review several recent articles and examine their experimental design. Classify them according to the types of design given here. Can you think of a way the design could be strengthened? If so, think of the potential costs of incorporating the design change. Could the project have been done that way with the resources available and clinical constraints?

2. Return to your experimental question. List several possible experimental designs. Critique each one in terms of strength of design and difficulty in execution.

3. If it is not clear to you that you have selected the strongest possible design to answer your question given your resources, relax! You can either consult a statistician or get additional training in experimental design. Although this article can raise awareness of design issues, it will not make you an instant expert. However, because the design is a critical feature of your project that cannot be changed later, it is essential to be certain of this before you start.

ACKNOWLEDGMENTS

We, as well as the editor of the Journal, extend our appreciation to the authors, who in the interest of improving research, were willing to have their work critiqued in print.

Reference