The Regression-Discontinuity Design

William M.K. Trochim, Ph.D.

Introduction

It is unfortunate that the regression-discontinuity design is so named. In everyday language both parts of the term have connotations that are primarily negative. To most people "regression" implies a move backwards or a return to some earlier, more primitive state, while "discontinuity" suggests an unnatural jump or shift in what might otherwise be a smoother, more continuous process. To a research methodologist, however, the term "regression-discontinuity" (RD) carries no such negative meaning. Instead, the RD design is seen as a useful method for determining whether or not a program or treatment is effective.

The label "RD design" actually refers to a set of design variations. In its simplest, most traditional form, the RD design is a pretest-posttest program-comparison group strategy. The unique characteristic that sets RD designs apart from other pre-post group designs is the method by which research participants are assigned to conditions. In RD designs, participants are assigned to program or comparison groups solely on the basis of a cutoff score on a preprogram measure. Thus the RD design is distinguished from randomized experiments (or randomized clinical trials) and from other quasi-experimental strategies by its unique method of assignment. This cutoff criterion implies the major advantage of RD design–it is appropriate for targeting a program or treatment to those who most need or deserve it. Thus, unlike its randomized or quasi-experimental alternatives, the RD design does not require that potentially needy individuals be assigned to a no-program comparison group in order to evaluate the effectiveness of a program.

The RD design has not been used frequently in social research. Most commonly, it has been implemented in compensatory education evaluation, where school children who obtain scores that fall below some predetermined cutoff value on an achievement test are assigned to remedial training designed to improve their performance. The low frequency of use may be attributable to several factors. Certainly, the design is a relative late comer. It was first devised in 1958 (Trochim, 1984) and initially discussed in the seminal work on quasi-experimental research design by Campbell and Stanley (1963, 1966). Its first major field tests did not occur until the mid-1970s, when it was incorporated into the nationwide evaluation system for compensatory education programs funded under Title I of the Elementary and Secondary Education Act (ESEA) of 1965 (Tallmadge and Wood, 1978).

In many situations, the design has not been used because one or more key criteria could not be met. For instance, RD designs force administrators to assign participants to conditions solely on the basis of quantitative indicators, thereby restricting the degree to which judgment, discretion, or favoritism may be used. Perhaps the most telling reason for the lack of wider adoption of the RD design is that at first glance the design doesn't seem to make sense. In most research, comparison (control) groups are used that are equivalent to program groups on preprogram indicators so that postprogram differences may be attributed to the program itself. Because of the cutoff criterion in RD designs, program and comparison groups are deliberately and maximally different on preprogram characteristics, an apparently insensible anomaly. An understanding of how the design actually works depends on at least a conceptual familiarity with regression analysis, thereby making the strategy a difficult one to convey to nonstatistical audiences.

Despite its lack of use, the RD design has great potential for evaluation and program research. From a methodological point of view, inferences that are drawn from a well-implemented RD design are comparable in internal validity to conclusions from randomized experiments. Thus the RD design is a strong competitor to randomized designs when causal hypotheses are being investigated. From an ethical perspective, RD designs are compatible with the goal of getting the program to those most in need. It is not necessary to deny the program to potentially deserving recipients simply for the sake of a scientific test. From an administrative viewpoint, the RD design is often directly usable with existing measurement efforts, such as the regularly collected statistical information typical of most management in-

Dr. Trochim is Associate Professor in Program Evaluation Studies in the College of Human Ecology at Cornell University.

formation systems. The advantages of the RD design warrant greater instructional efforts by the methodological community to encourage its use where appropriate.

The major focus of this article is to discuss the applicability of the RD design to health evaluation. The design seems especially suitable for many health contexts because of the abundance of quantitative indicators and information data bases, as well as the trend toward greater accountability through the use of specific quantitative variables as the basis for allocating health resources. In addition to the traditional reliance on quantitative indicators in medicine (e.g., blood pressure readings, temperature, pulse, severity of symptomatology) and psychiatry (e.g., MMPI scores, DSM classifications), there is increasing reliance on quantitative information in the Medicaid and Medicare reimbursement systems, in the movement toward greater quality and efficiency of hospital care, in the development of the Diagnostic-Related Groups (DRG) classification systems, in the nursing home quality of care movement, and in many other health fields. To the extent that these quantitative indicators are used or can be used in connection with cutoff values for determining eligibility for programs, treatments, or other resources, the RD design may be the most appropriate and feasible strategy for evaluation.

This article will describe the basic RD design and examine the variations that may apply to health contexts. A major goal is to provide health administrators and evaluators with enough information to judge intelligently whether it is feasible to use an RD design to assess a specific evaluation problem. A comprehensive discussion of the statistical analysis of the RD design is outside the scope of this work, but a general analytic model will be presented, and the major analytic issues will be discussed. Because RD designs have not been applied formally in health contexts, the statistical analysis is illustrated using data taken from an evaluation of a compensatory educational reading program, but the procedures should be directly generalizable to data collected in health fields.

The Logic and Structure of RD Designs

The basic RD design. The term "basic RD design" refers to the design as it was originally conceived and discussed. In Campbell and Stanley (1963, 1966), the RD design was presented as a pretest-posttest two-group design. The term "pretest-posttest" implies that the same measure (or perhaps alternate forms of the same measure) is administered before and after some program or treatment. (In fact, the RD design does not require that the pre-and posttest measures be the same.)

Throughout this article, the term "pretest measure" will be used to imply that the same measure is given twice, while the term "preprogram measure" will imply more broadly that before-and-after measures may be the same or different. It is assumed that a cutoff value on the pretest or preprogram measure is being used to assign persons or other units to the program. Two-group versions of the RD design might imply either that some treatment or program is being contrasted with a no-program condition or that two alternative programs are being compared. The description of the basic design as a two-group design implies that a single pretest cutoff score is used to assign participants to either the program or comparison group. Here, the term "participants" will refer to whichever unit is assigned. In many cases participants are individuals, but they could be any definable unit, such as hospital wards, hospitals, or counties. The term "program" will be used throughout to refer to any program, treatment, or manipulation for which effects are being examined. In notational form, the basic RD design might be depicted as

```
C O X O

C O   O
```

where the C indicates that groups are assigned by means of a cutoff score, an O stands for the administration of a measure to a group, an X depicts the implementation of a program, and each group is described on a single line (i.e., program group on top, control group on the bottom).

An example will make this initial presentation more concrete: for this purpose, a hypothetical study will be described where the interest is in examining the effect of a new treatment protocol for inpatients with a particular diagnosis. For simplicity, it can be assumed that the new protocol will be tried on patients who are considered most ill. For each patient there is a continuous quantitative indicator of severity of illness that is a composite rating that can take values from 1 to 100, where high scores indicate greater illness. Furthermore, a pretest cutoff score of 50 was (more or less arbitrarily) chosen as the assignment criterion, and all those scoring 50 or higher on the pretest are to be given the new treatment protocol, while those with scores lower than 50 are given the standard treatment.

It is useful to begin by considering what the data might look like if the treatment protocol was not administered, but instead all participants were measured at two points in time. Figure 1 shows the hypothetical bivariate distribution for this situation. Each dot on the figure indicates a single person's pretest and posttest scores. The dot labeled "a" shows an individual who had a high pretest and posttest score; this person was severely ill on the first measure and remained so on the second. The dot labeled "b" represents the pretest and posttest for an

Figure 1. Hypothetical no-program basic RD design

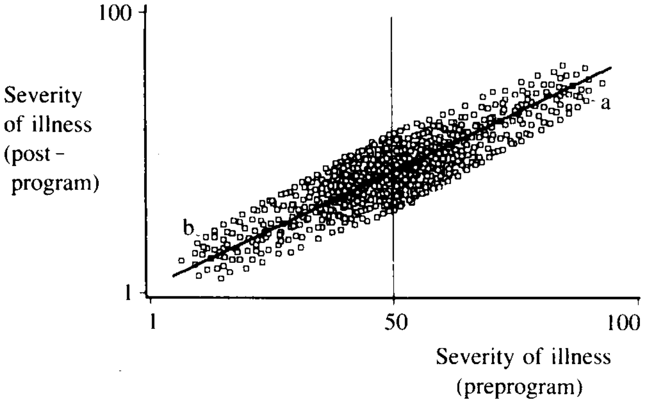
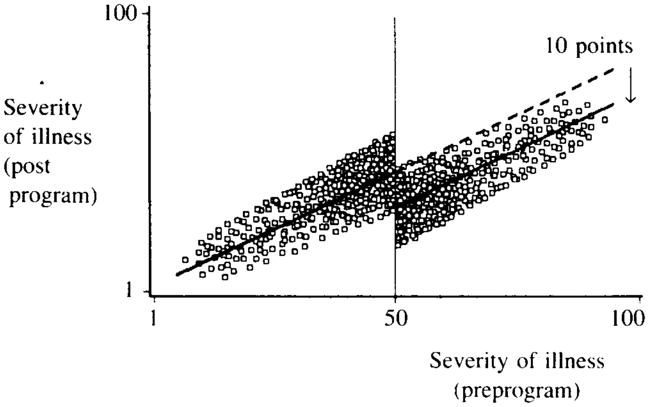


Figure 2. Hypothetical RD design with a constant (additive) program effect



individual who was not severely ill on both occasions. The vertical line at the pretest score of 50 indicates the cutoff point (although for Fig. 1 it is assumed that no treatment protocol has been given). The solid line through the bivariate distribution is the linear regression line. The distribution depicts a strong positive relationship between the pretest and posttest–in general, the more severely ill a person is at one point in time, the more ill that person will be at the other.

Figure 2 illustrates what the outcome might look like if the new treatment protocol is administered and has a positive effect. For simplicity, it will be assumed that the treatment protocol had a constant effect that lowered each treated person's severity of illness by 10 points. Figure 2 is identical to Figure 1 except that all points to the right of the cutoff (i.e., the new treatment protocol group) have been lowered by 10 points on the posttest. The dashed line in Figure 2 shows what the treated group's regression line would be expected to look like if the program had no effect (as was the case in Fig. 1). On the basis of Figure 2, it can be seen how the RD design got its name–a program effect is implied when a "discontinuity" in the "regression" lines is observed at the cutoff point.

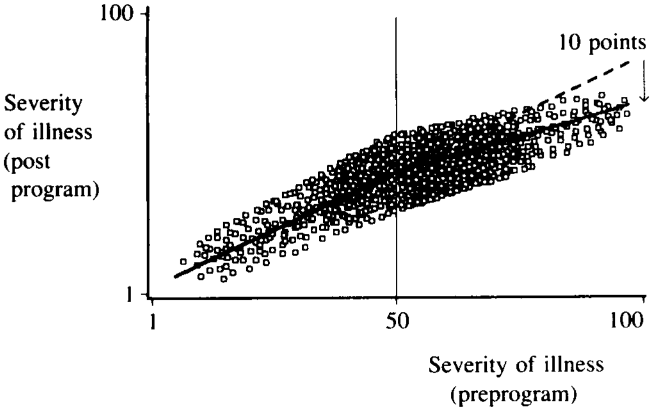
Figure 2 portrays a very simplistic version of the design with an unrealistically uniform outcome. This could be complicated by assuming that instead of a constant effect, the program had no effect on persons scoring at the cutoff and its greatest effect (again, 10 points) on those most severely ill. This hypothetical program-pretest interaction effect case is shown in Figure 3. Again, the dashed line shows the regression line that would be expected if the treatment was not effective. As in all RD designs, it is the discontinuity (in this case, the change in slopes) in the regression lines at the cutoff point that implies the treatment has an effect.

As in any research, there is a wide variety of possible outcomes. Instead of being effective, a program might actually harm the participants, as evidenced on outcome measures. If such a negative effect occurs and is constant across the pretest range, it displaces the treatment group regression line in Figure 2 upward instead of downward. A similar argument can be extended for the case of a negative interaction effect.

Selection of the cutoff. The choice of cutoff value is usually based on one of two factors. It can be made solely on the basis of the program resources that are available. For instance, if a program has the capability of handling only 25 persons and 70 people apply, a cutoff point can be chosen that distinguishes the 25 most needy applicants from the rest. Alternatively, the cutoff can be chosen on substantive ground. If the preprogram assignment measure is an indication of severity of illness measured on a scale of 1 to 7 and physicians or other experts believe that all patients scoring 5 or more are critical and are a "good" fit with the criteria defined for program participants, then a cutoff value of 5 may be used.

Interpretation of results. In order to interpret the results of an RD design, the nature of the assignment variable must be known, as well as who received the program and the nature of the outcome measure. With-

Figure 3. Hypothetical basic Rd design with a program-pretest interaction effect



out this information, there is no distinct outcome pattern that directly indicates whether an effect is positive or negative.

To illustrate this, it will be assumed that a hospital administrator would like to improve the quality of patient care by implementing an intensive quality of care training program for staff. Because of financial constraints, the program is too costly to implement for all employees; instead, it will be administered to the staff of specifically targeted units or wards that seem most in need of improvements in quality of care. Two general measures of quality of care are available. The first is an aggregate rating of quality of care based on observation and rating by an administrative staff member–labeled here as the "QOC rating." The second is the ratio of the number of recorded patient complaints relative to the number of patients in the unit over a fixed period of time, termed here the "complaint ratio." In this scenario, the administrator could use either the QOC rating or the complaint ratio as the basis for assigning units to receive the training. Similarly, the effects of the training could be measured on either variable.

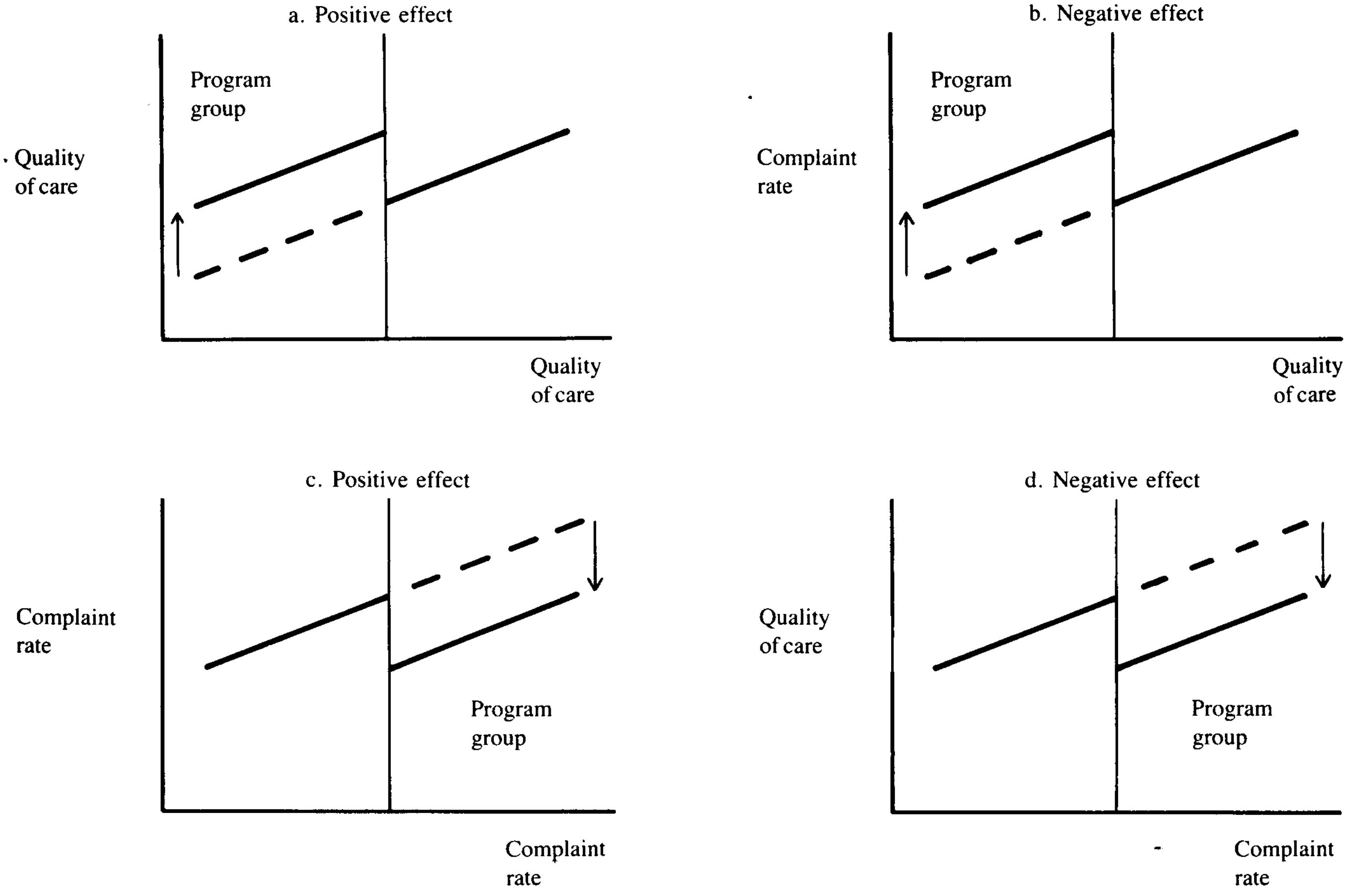
Figure 4 shows four outcomes of alternative RD implementations possible under this scenario; only the regression lines are shown. It is worth noting that even though all four outcomes have the same pattern of regression lines, they do not imply the same result. In Figures 4a and 4b, hospital units were assigned to training because they scored "below" some cutoff score on the QOC rating. In Figures 4c and 4d, units were given training because they scored "above" the cutoff score on the complaint ratio measure. In each instance, the dashed line indicates the regression line that would be expected for the training group if the training had no effect. This dashed line represents the no-discontinuity projection of the comparison group regression line into the region of the program group pretest scores.

Clearly, even though the outcome regression lines are the same in all four groups, the four graphs would be interpreted differently. Figure 4a depicts a positive effect because training raised the program group regression line on the QOC rating over what would have been expected. However, Figure 4b shows a negative effect because the program raised training group scores on the complaint ratio, indicating increased complaint rates. In Figure 4c, a positive effect is seen, because the regression line has been lowered on the complaint ratio relative to what could have been expected. Finally, Figure 4d shows a negative effect where the training resulted in lower QOC ratings than would be expected otherwise. The point here is a simple one. A discontinuity in regression lines indicates a program effect in the RD design, but the discontinuity alone is not sufficient to determine whether the effect is positive or negative. In order to make this determination, it is necessary to know who received the program and how to interpret the direction of scale values on the outcome measures.

The role of the comparison group in RD designs. The purpose of the foregoing discussion has been to explain the benchmark for comparison in the RD design. In experimental or other quasi-experimental designs, researchers either assume or try to provide evidence that the program and comparison groups were equivalent prior to the program so that postprogram differences can be attributed to the manipulation. The RD design involves no such assumption. Instead, with RD designs it is assumed that–in the absence of the program–the pre-post relationship would be equivalent for the two groups. Thus, the strength of the RD design is dependent on two major factors. The first is the assumption that there is no spurious discontinuity in the pre-post relationship that happens to coincide with the cutoff point. The second factor concerns the degree to which the pre-post relationship is known and correctly modeled; this constitutes the major problem in the statistical analysis of the RD design.

The RD design from a pattern matching perspective. Another way to understand how the RD design works is to reformulate it in terms of a pattern matching philosophy of research (Trochim, 1985). The general pattern matching notion is that in all confirmatory, hypothesis-testing research there is an ideal or theoretical pattern and a real or obtained pattern. The theoretical pattern describes the researcher's expectation for the results of the study; the obtained pattern consists of the

Figure 4. Interpretations of direction of effects for four hypothetical RD designs



results as measured or observed. The theory gains support only if the theoretical and obtained patterns match, and there are no alternative plausible theoretical patterns that would match the obtained outcome as well. The central principle of pattern matching is that the more distinctive, unique, or idiosyncratic the theoretical pattern, the more likely it is that a match will support the theory, because in general it will be less likely that there will be plausible alternative theoretical patterns to the one being tested.

With this in mind, the strengths of the RD design can be better understood, especially in comparison to other quasi-experimental strategies. In the RD design the theoretical pattern is the expectation of the discontinuity in regression lines at the cutoff point. It is very unlikely that a cutoff point would be chosen that just happened to coincide with some naturally occurring discontinuity in the pre-post relationship. In fact, typical pre-post relationships seldom evidence such natural discontinuities at all. Thus the selection of the cutoff usually constitutes the creation of a relatively unique theoretical expectation that, if confirmed in the obtained pattern, will allow few plausible counter explanations.

This can be contrasted with the typical pretest-posttest nonequivalent group design. Here, groups are nonrandomly assigned in the hope that they are equivalent on preprogram characteristics. If there is any preprogram characteristic on which the groups differ that is related to outcome measures, a difference between groups on an outcome measure may be attributed to this characteristic rather than to the program. In most research settings, there usually are an abundance of such preprogram group differences that are plausible or at least possible. In the RD design, it is seldom plausible that the groups would be expected to differ naturally in their pre-post relationship at a point that coincides with the cutoff. Thus it is the dichotomization of the bivariate distribution at the chosen cutoff point that determines the unique theoretical pattern that provides the strength of the RD design from a pattern matching perspective.

RD designs and internal validity. Over the past two decades the theory of internal validity has been formulated and articulated largely through the work of Donald T. Campbell (Campbell and Stanley, 1963, 1966; Cook and Campbell, 1979). More recently, Campbell (1986) has relabeled the concept of internal validity as "local molar causal validity." Because of the tentativeness of the relabeling, this discussion will use the more traditional term. "Internal validity" refers to whether or not it can be inferred that the treatment or program being investigated caused a change in outcome indicators. Internal validity is not concerned with the ability to generalize; rather, it focuses on whether a causal relationship can be demonstrated for the immediate research context. Research designs that address causal questions are often compared on their relative ability to yield internally valid results.

In most causal hypothesis tests, the central inferential question is whether any observed outcome differences between groups are attributable to the program or instead to some other factor. In order to argue for the internal validity of an inference, the analyst must attempt to demonstrate that the program–and not some plausible alternative explanation–is responsible for the effect. In the literature on internal validity, these plausible alternative explanations or factors are often termed "threats" to internal validity. A number of typical threats to internal validity have been identified. For instance, in a one-group pre-post study a gain from pretest to posttest may be attributable to the program or to other plausible factors, such as historical events occurring between pretest and posttest or natural maturation over time.

Many threats can be ruled out with the inclusion of a control group. Assuming that the control group is equivalent to the program group prior to the study, the control group pre-post gain will provide evidence for the change that should be attributed to all factors other than the program. A different rate of gain in the program group provides evidence for the relative effect of the program itself. Thus, randomized experimental designs are considered strong in internal validity because of confidence in the probabilistic preprogram equivalence between groups that results from random assignment and helps assure that the control group will provide a legitimate reflection of all nonprogram factors that might affect outcomes.

In designs that do not use random assignment, the central internal validity concern is the possibility that groups may not be equivalent prior to the program. The term "selection bias" refers to the case where preprogram differences between groups are responsible for postprogram differences. Any nonprogram factor that is differentially present across groups can constitute a selection bias or a selection threat to internal validity.

Because of the deliberate preprogram differences between groups in RD designs, there are several selection threats to internal validity that might appear to be a problem. For instance, a selection-maturation threat implies that different rates of maturation between groups might explain outcome differences. For example, a pre-post distribution with a linear relationship having a slope equal to two units implies that, on the average, a person with a given pretest score will have a posttest score two times as high. Clearly there is maturation in this situation–that is, subjects are achieving consistently higher scores over time. For a subject with a pretest score of 10 units, a posttest score of 20 would be predicted for an absolute gain of 10. But, if the person has a pretest score of 50, the prediction would be a posttest score of 100, for an absolute gain of 50. Thus the second person naturally gains or matures more in absolute units (although the rate of gain relative to the pretest score is constant). Along these lines, in the RD design it is expected that all participants may mature, and in absolute terms, this maturation may be different for the two groups on average. Nevertheless, a program effect in the RD design is not indicated by a difference between the posttest averages of the groups, but rather by a change in the pre-post relationship at the cutoff point. In this example, although different absolute levels of maturation are expected, a single "continuous" regression line with a slope equal to 2 would describe these different maturational rates. In order for selection-maturation to be a threat to internal validity in RD designs, it must induce a discontinuity in the pre-post relationship that happens to coincide with the cutoff point–an unlikely scenario in most studies.

Another selection threat to internal validity that might intuitively seem likely is the possibility of differential regression to the mean or a selection-regression threat. The phenomenon of regression to the mean arises when there is an asymmetrical sampling of groups from a distribution. On any subsequent measure the obtained sample group mean will be closer to the population mean for that measure (in standardized units) than the sample mean from the original distribution is to its population mean.

In RD designs, asymmetric samples are deliberately created, and consequently, regression toward the mean is expected in both groups. In general, the low-scoring pretest group is expected to evidence a relative gain on the posttest, and the high-scoring pretest group is expected to show a relative loss. As with selection-maturation, even though differential regression to the mean is expected, it poses no problem for the internal validity of the RD design. It is not expected that regression to the mean will result in a discontinuity in the bivariate relationship coincidental with the cutoff point. In fact, regression to the mean is expected to be continuous across the range of the pretest scores and to be described by the regression line itself. [Draper and Smith (1981) point out that the term "regression" was originally used by Galton to refer to the fact that a regression line describes regression to the mean.]

Although initially the RD design may seem susceptible to selection biases, it is not. The above discussion demonstrates that only factors that would naturally induce a discontinuity in the pre-post relationship could be considered threats to the internal validity of inferences from the RD design. In principle, the RD design is as strong in internal validity as its randomized experimental alternatives. However, in practice the validity of the RD design depends directly on how well the analyst can model the true pre-post relationship, certainly a nontrivial statistical problem.

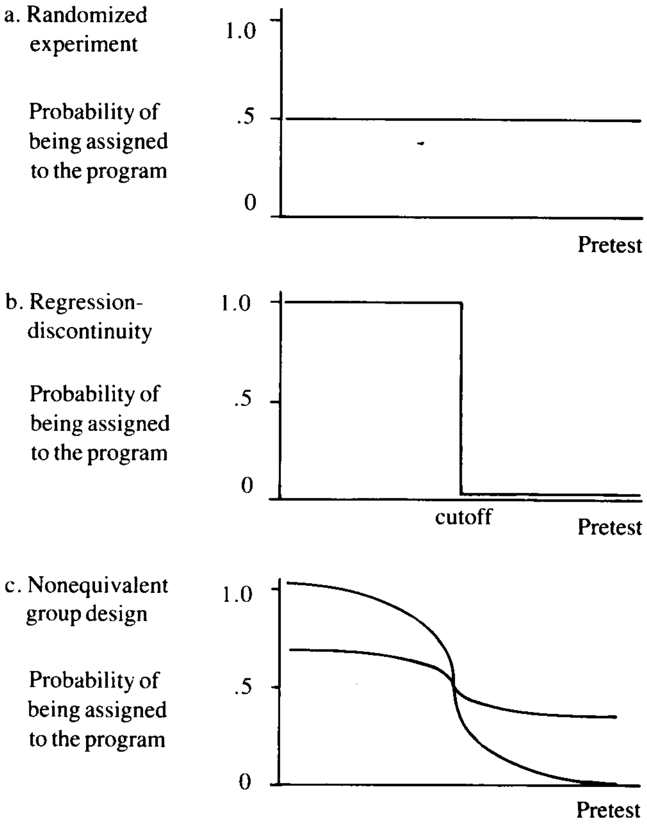
Relationship of RD designs to other designs. The previous discussion implies how RD designs might be viewed relative to their closest design alternatives. The characteristic that differentiates pretest-posttest group designs is the method of assignment of participants to groups. Figure 5 shows the probability of being assigned to the program group, given the pretest score for the three major types of pretest-posttest group designs. Figure 5a shows that in the simple randomized experiment or randomized clinical trial the probability of being assigned to the program group is .5 regardless of pretest score, because participants are assigned randomly. Figure 5b shows that in an RD design where persons below the cutoff are assigned to the program, the probability of assignment is a step function–those below the cutoff have a probability of 1.0 and those above have a zero probability of being assigned to the program.

Randomized experiments and RD designs are extreme cases in terms of the conditional probability of assignment to the program, given the pretest score. All nonequivalent group designs fall on a continuum somewhere between these extremes. Figure 5c shows assignment functions for two hypothetical nonequivalent group designs. Both functions show that there is pretest nonequivalence between groups, with the program group scoring lower on the pretest on average. In one case, the groups are nearly equivalent with a slightly greater tendency for low pretest scores to be assigned to the program group. This function more closely approximates the randomized experimental one and shows that the groups are nearly equivalent. The other case shows more marked nonequivalence on the pretest, more closely approximating the RD function of Figure 5b.

The key is that for both randomized experiments and RD designs the analyst knows exactly what the true assignment function is because it is dictated by the nature of the designs. With nonequivalent group designs, this assignment function is never known perfectly and must be estimated. Thus randomized experiments and RD designs are strong against selection bias because the rule for selection (assignment to groups) is known perfectly.

This discussion could be recast in terms of the advice that methodologists might give to policymakers and administrators on how to improve the accountability of their programs. If the goal is to assess the effects of a

Figure 5. Comparison of pre-post design assignment functions



program (i.e., to examine a causal hypothesis), both randomized experiments and RD designs are strong choices. It makes sense intuitively that the accountability of a program is largely dependent on the explicitness of the assignment or allocation of the program to recipients. Lawmakers and administrators need to recognize that programs are more easily evaluated and more accountable when the allocation of the program is more public and verifiable. The three designs described in Figure 5 are analogous to the three types of program allocation schemes that legislators or administrators might choose.

In randomized experiments, the assignment processs is analogous to a lottery, while RD designs can be considered explicit, accountable methods for assigning program recipients on the basis of need or merit. Nonequivalent group designs might be considered a type of political allocation because they enable the use of unverifiable, subjective, or politically motivated assignment (of course, most social programs are politically allocated). Even when programs are allocated primarily on the basis of need or merit, the regulatory agency usually reserves some discretionary capability in deciding who receives the program. Without debating the need for this, it is clear that those who seek program accountability should be encouraged to establish explicit criteria for program eligibility, either through probability based lotteries or by relying on quantitative eligibility ratings and cutoff values as in the RD design. To the extent that legislators and administrators move toward more explicit assignment criteria, both the potential utility of the RD design and the accountability of the programs will be increased.

RD Design Variations

If the RD design were limited only to the basic form described above, it would still have great utility as a method for evaluation in health. But the design has many variations that increase its versatility and utility. This section describes many of these variations and illustrates their application in hypothetical evaluation problems from health-related areas.

Assignment variations. While the use of a preprogram cutoff value is the distinguishing feature of the RD design, it often proves to be difficult to implement. When using the RD design in compensatory education, school districts routinely set up formal procedures for violating the assignment by cutoff rule. More often than not, such procedures are couched in terms that emphasize the educational and political advantages and minimize the methodological difficulties that are introduced. For instance, several school districts have used a "challenge program" that allowed teachers, parents, or administrators to appeal the decision based on the cutoff score, usually on educational grounds. A teacher could "challenge" a student into the program because in the teacher's professional judgment the student needed the program, even though the student may have scored above the cutoff value on the preprogram achievement test. Similarly, students could be challenged out of the program even though they scored below the cutoff if teachers or parents felt that the test score was inaccurately low or that participation in the program might be stigmatizing or otherwise deleterious.

There is good reason to believe that comparable pressures to make exceptions to cutoff-based assignment would arise in health contexts. For instance, is it realistic to envision the automatic assignment of patients to a treatment program if the judgment of the attending physician contradicts the assignment by cutoff? Similarly, is it possible that political favoritism, scheduling factors, and other anomalies won't affect the assignment of hospital staff members to in-service quality-control training sessions regardless of their status relative to the cutoff? Except in the unlikely case where exceptions to the cutoff rule are random, misassignment constitutes a major potential bias in RD designs because it will usually introduce discontinuities in the pre-post relationship at the cutoff or complicate the form of the pre-post distribution. Trochim (1984) used statistical simulations to show that under reasonable assumptions about challenges in compensatory education, the resulting bias could be significant.

The major reason for exceptions to the cutoff rule is that it is perceived as too restrictive or inflexible. At first glance, it appears that the cutoff criterion does not allow room for professional judgment or discretion. In fact, the RD design doesn't preclude such judgment–it only requires that judgments be quantified or explicitly accounted for in the selection of the cutoff. To explain this, two major strategies for including more discretion in assignment are discussed: the use of multiple cutoff points and the use of multiple preprogram assignment variables.

Multiple cutoff points. RD designs are not limited to the use of a single cutoff value. It is assumed that the pre-post relationship in the absence of the treatment can be described with a single continuous regression line that extends over the entire range of the preprogram scores. Even when multiple cutoff points are used, the comparison group's pre-post relationship would still be used as the benchmark for projecting where other groups' regression lines should be.

The simplest case would involve the use of two cutoff scores. Persons scoring above the higher cutoff would be assigned to one group, those scoring below the lower cutoff would be assigned to the other, and those scoring between the cutoffs (or within the cutoff interval) could be assigned by a variety of methods. To illustrate this, consider a hypothetical case where the attending psychiatrist is to assign patients with a specific diagnosis to receive either a novel intensive treatment protocol or the traditional therapy for the type of disorder. The preprogram assignment measure is a 1-7 severity rating for the psychosocial stressors as described in the *Diagnostic and Statistical Manual of Mental Disorders* (American Psychiatric Association, 1980). This will be rated by the psychiatrist and key nursing staff members, and the average for each person will be used for assignment. The descriptive terms for scale values on the "severity of psychosocial stressor" measure indicate that values from 1 to 3 imply from "none" to "mild" stress, a 4 implies "moderate" stress, and a 5 or greater indicates "severe" stress or worse. The assumption is that high scorers are under more severe psychosocial stress and are more in need of the intensive treatment.

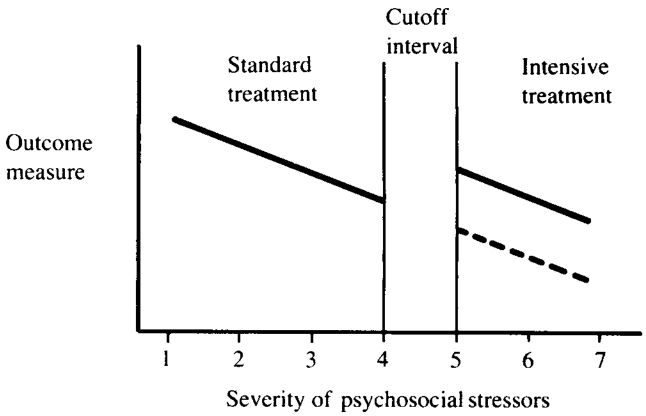
Nevertheless, it is reasonable to expect that the psychiatrist may not have absolute faith in the validity of this single quantitative scale. The original plan was to assign all patients scoring 5 or higher on this averaged rating to the new program and those scoring lower than 5 to the traditional treatment. But the psychiatrist is concerned that some of those who score just below 5 may be under more stress than the measure indicates, and further, believes that professional judgment can be used to distinguish cases in need. The revised plan utilizes two cutoff values. All those who score 5 or greater are automatically assigned to the new program, those scoring less than 4 are automatically assigned to standard treatment, and those who score in the cutoff interval are assigned entirely on the basis of the professional judgment of the psychiatrist. This multiple-cutoff RD design is shown in Figure 6. In the figure, it is assumed that the outcome indicator is some measure of performance, with higher scores indicating relative improvement. Here it can be seen that the program had a positive effect on the outcome.

Several points are noteworthy in this hypothetical case. A basic requirement of the RD design is enough information for the analyst to model the comparison group pre-post relationship. If the preprogram assignment measure had been rated by only one person (e.g., the psychiatrist alone), the resulting value would be an integer between 1 and 7; for the comparison group, preprogram scores would take only the integer values 1, 2, or 3. Only these three values would be available for modeling the comparison group regression line–a marginally sufficient degree of variability if the true relationship is linear and an insufficient degree if curvilinearity is present. By averaging the ratings of a number of individuals, greater discrimination is introduced into the measure, thus increasing the ability to estimate the true pre-post relationship.

In the example, subjects within the cutoff interval are assigned solely on discretionary grounds. Consequently, multiple analyses of the data are warranted, beginning with an estimation of the program effect excluding all cases that fall within the interval as shown in Figure 6. Subsequently, it would be reasonable to include all cases to see if the original estimate is different. Greater credibility would be accorded to the former analysis than to the latter because of the greater potential for selection biases in the discretionary cases.

The use of multiple cutoff points can enhance the internal validity of a study if random assignment is used on cases that fall within the cutoff interval. This might arise in practice if the original design required random assignment of all participants, but such assignment was deemed unethical because it would deny the most needy a potentially beneficial treatment. Here, the RD design is "coupled" with a randomized experiment that is limited to a restricted range of the pretest falling within the cutoff interval as first suggested by Boruch (1975). One advantage of this approach (outside of the obvious ethical issue) is that the statistical power of the resulting analysis, which would use only the cases within the interval, would be treated as a traditional randomized experiment; one would use only the scores outside the interval and be treated as a basic RD design; and one would use all scores resulting in an analysis with greater statistical power than the other two.

Figure 6. Example of an RD design with multiple cutoffs



The use of multiple cutoff points also could be useful in institutional settings where a serial implementation of the program could be done, beginning with those most in need and progressing to those who are less needy. For example, a consortium of hospitals might seek to reduce the incidence of lower back injuries among their nursing staff by providing training workshops for hospital units where tasks are performed that are likely to cause such injury. The training sessions cannot be provided to all units at one time, but eventually all units should be trained. The hospitals collect data on the incidence rates of lower back injuries at regular intervals. A strategy for accomplishing this within an RD framework is illustrated in Figure 7.

The figure shows three waves of implementation: in the first wave (a), all units scoring above the cutoff are assigned to training workshops; in the second wave (b), a second and lower cutoff is used; in the third wave (c), another, even lower cutoff is chosen. It is important to note that for all three waves the original preprogram measure at time 1 is used for assignment. The posttest is lower back injury rates, and these are measured on all units after each wave of implementation. In design notation this can be depicted as follows:

C O X_1 O O O

C O O X_2 O O

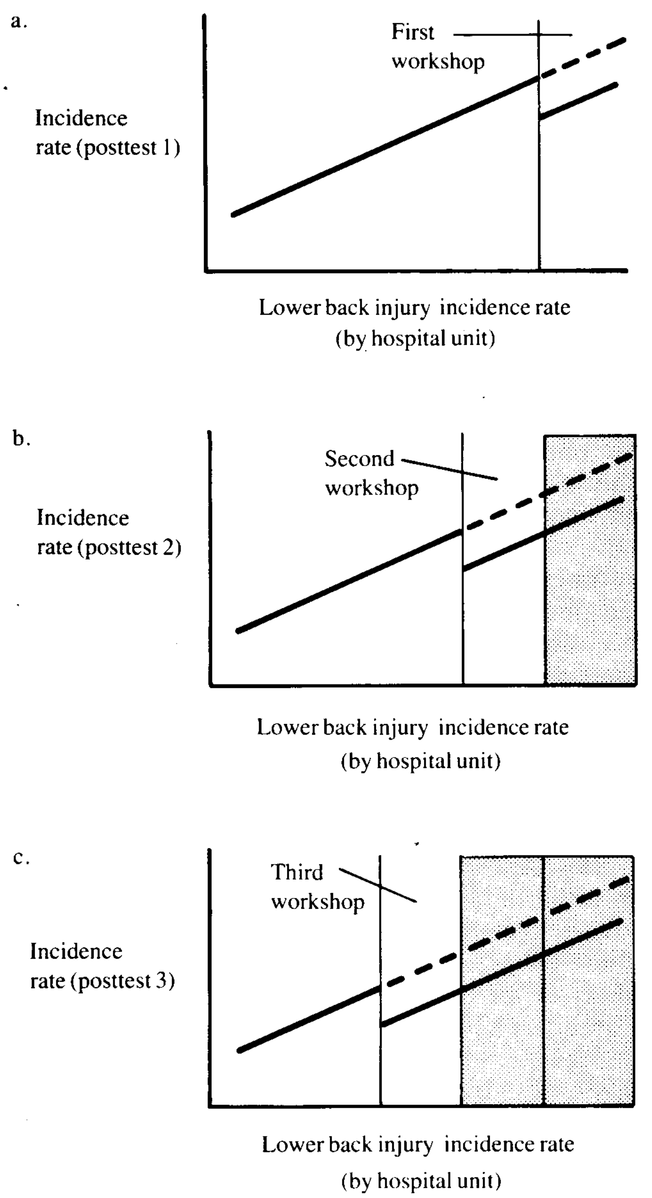
C O O O X_3 O

C O O O O

where the subscripts of X indicate the wave of training workshop implemented. Again, multiple analyses are called for, with the control group incorporating a smaller pretest range on each successive analysis. A strategy of this type may be used whenever an institution has a regularly instituted program, and it is feasible to collect data on all participants at each wave of the study. The design meets the institutional desire to treat the most needy first and has the methodological advantages of providing replication of the program evaluation and enabling easy coupling in situations where data are routinely collected by the institution at regular intervals.

Multiple assignment measures. In the basic RD design the requirement of strict adherence to the cutoff criterion may be unreasonable because it relies on a single quantitative indicator that may not capture well the degree of preprogram need. Here, two strategies for incorporating multiple preprogram measures are discussed: the use of separate measures, each having its own cutoff value, and the use of an aggregate index variable that is a composite of several measures.

Figure 7. RD designs with sequential programs



The example used to illustrate these will concern the diagnostic difficulties that face the clinician who must decide the severity of illness as the basis for treatment assignment. In this example, the physician will use the criteria for a major depressive episode with melancholia as described in the *Diagnostic and Statistical Manual of Mental Disorders* (American Psychiatric Association, 1980).

Six major variables are used to make such a diagnosis and presumably to judge its severity. Three of these are exclusionary criteria (e.g., no evidence of schizophrenia) that would be used to exclude cases from the study on grounds that they do not fall within the diagnosis. The remaining three variables are ratings of symptoms (e.g., dysphoric mood, symptoms of depression, and melancholia symptoms) that must be present for varying periods of time; these are the measures that would be used to make the diagnosis and as the basis for assignment to treatment for the disorder.

For example, consider the case where these three measures would be used separately, with each having its own cutoff value. In this example, it is assumed that a patient will be assigned to the treatment if that patient receives a score greater than 4 on dysphoric mood, has at least four of the eight depressed symptoms, and at least three of the six melancholia symptoms. If any of these criteria are not met, the patient will be in the control group. Here, all three measures also will be used as posttests. Thus there are nine (i.e., three pre × three post) bivariate distributions in this case. None of the bivariate distributions is likely to adhere strictly to the cutoff criterion because there are likely to be some cases that meet one or two of the cutoff rules but not all three. Nevertheless, the multivariate conditions are satisfied because all three cutoffs must be met for a patient to be placed in the treatment.

This approach would not work if the cutoff criteria were joined by "or" instead of "and." If the assignment rule was that the patient had to meet any two of the three cutoff rules, then a selection bias would be likely.

However, an "or" rule can be used when there is a relatively large set of assignment variables. This can be illustrated with an example involving ten or more variables for assignment–each with its own cutoff–some of which are more objective symptomatically related ratings or measures (e.g., blood pressure, severity of illness), while others are more subjectively based ratings (by physicians or nursing staff) of need for treatment. Further, the cutoff criterion is that the cutoffs must be met on six or more of any of the measures for a patient to be assigned to the program. Here the assignment measure is actually a count of the number of individual assignment criteria that the person meets. With 10 assignment measures a person could obtain a score between 0 (exceeds cutoff on no measure) and 10 (exceeds cutoff on all ten measures). Thus the real assignment measure in this example is based on an "or" condition that takes into account a large set of individual measures. A relatively large set of assignment measures is necessary in this design to have sufficient preprogram variability for estimating a comparison group regression line. This variation would not work if there were three assignment measures (any two of which must be met) because the comparison group preprogram values would consist of the two integers 0 (no criteria met) and 1 (only one criterion met). Nevertheless, in situations where many assignment measures are available, this RD variation may be particularly useful.

Neither of these multiple-measure versions of the RD design has ever been suggested prior to this writing (as far as this author knows). Both versions have obvious value, in that they allow for the use of multiple quantitative assignment variables, each with its own cutoff. If physicians or hospital administrators are reluctant to rely on a single quantitative assignment measure or if they wish to take the professional judgment of the physician or staff into account, any number of additional ratings can be included explicitly within the design.

Another variation of the RD design that utilizes multiple preprogram assignment variables occurs when the individual variables are combined into a single index variable on which a cutoff is selected that constitutes the sole basis for assignment to group. The major difficulty with this procedure concerns how the variables should be combined. In the simplest case, the variables can be added or averaged. However, this would not be appropriate if the variables are on different scales or have different means and standard deviations.

Standardizing each variable before aggregation will almost always be desirable. In most cases, it also will be desirable to weight certain variables as more important than others. In the example given above, it may not be desirable just to add or average the three standardized variables of dysphoric mood, depressed symptoms, and melancholia symptoms. Instead, the first two indicators could be weighted more heavily than the third. Weights could be based on theory or on prior empirical investigation of the variables. For an aggregate index it may be more difficult to arrive at a theoretically sensible cutoff value because the meaning of a number on the index is related to need in a more complex fashion and may be difficult to interpret.

It is possible that these approaches to the use of multiple preprogram assignment variables can be combined. Some subsets of variables could be pooled into aggregate indexes, while others would be left as they are. Each variable or index of variables could have its own cutoff value. These multiple indicators could be combined with either an "and" or an "or" rule as described above. However it is accomplished, the use of multiple preprogram assignment variables enables the researcher to develop more sensitive and accurate assignment rules and should help to reduce resistance to the adherence to the cutoff criterion in RD designs.

Program variations. The basic RD design assumes that the comparison group is actually a no-program control. This gives the appearance of an "absolute" contrast, that is, the program versus nothing at all. Essentially, the research question is whether or not the program makes a difference when compared to doing nothing. In many health contexts, this type of absolute question may be impractical, unethical, or meaningless. For instance, there probably would be more interest in examining whether the treatment has an effect relative to the current standard treatment for the disorder. Or, there may be several potentially useful and novel treatments that should be tested. In other situations, the interactions of different treatment protocols could be examined. All of these variations are possible with RD designs, but considerable thought must be given to how they will be incorporated.

Clearly, any two treatments (whether an absolute or relative comparison) may be incorporated within the basic RD design structure. However, one of the major advantages of the RD design is that persons who are in greater need may be assigned to progressively riskier treatments. For example, if there are three treatment programs of interest–the standard, well-accepted treatment and two experimental and riskier protocols–and there is reason to believe that the first experimental treatment is riskier than the other, it may be desirable to have two cutoff points. In this scenario, the neediest patients would be assigned to the most risky program, the least needy patients would be assigned to the least risky treatment (i.e., standard treatment), and those who fall between the two cutoff points would be assigned to the moderately risky program. On the other hand, if there is no *a priori* basis for judging one experimental program riskier than the other, a single cutoff may be used to assign those who are least in need to standard treatment; those on the other side of the cutoff would be randomly assigned to one of the two experimental protocols. Even more complex assignment patterns can be envisioned, with multiple cutoffs and perhaps multiple groups within certain cutoff intervals; but these become less feasible as either within-group sample size or comparison group preprogram variability decreases. Nevertheless, it is worth recognizing that RD designs allow considerable flexibility in examining alternative program variations.

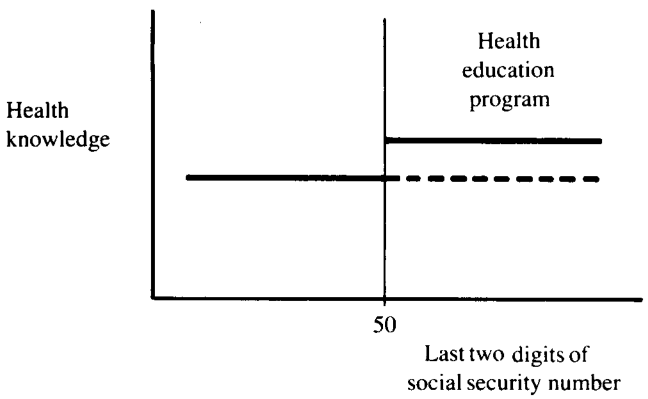
Postprogram measurement variations. The foregoing has shown that the basic RD design, which utilizes a single postprogram measure identical to the pretest, can be extended through variations of the assignment strategy and the program comparisons that are made. In this section, the measurement of postprogram variables will be considered, including the use of (1) posttests that differ from the assignment measure, (2) multiple posttests and subtests, and (3) posttests that are not continuous normal variables.

Several of the examples provided earlier implied that pre- and postprogram measures don't have to be the same or equivalent in RD designs. Initially, this may seem counter-intuitive, because usually in most pre-post designs it is important to include the same measure at both points in time or, less desirably, to use a proxy preprogram measure that purportedly taps the same construct. This is in part because often there is interest in pre-post gains and in part because a pretest may be the best covariate that can be included. In RD designs it is not necessary to use the same pre- and postprogram measures; at times, it may even be undesirable. For instance, subjects could be assigned to a health education program on the basis of economic need (e.g., income), but the results of that program could be examined on the basis of indicators of knowledge and attitudes toward health. Here, the assignment variable is the economic indicator, and the postprogram measures reflect the substantive interests of the program.

The following example illustrates why RD designs don't require pre-post isomorphism and explains the relationship between RD designs and randomized experiments. In this hypothetical case, subjects are assigned to a health education program on the basis of the last two digits of their Social Security numbers–if the value is greater than 50, they receive the program, otherwise they do not. The assumption here is that the assignment measure is statistically unrelated to knowledge–that is, the correlation is zero. The outcome measure is an objective test of their knowledge of health issues as taught in the program. If the program increases knowledge scores, regression analysis results like those shown in Figure 8 might be expected.

In the figure, the relationship between measures is assumed to be zero; therefore, the regression lines have zero slopes. As always, the program effect is indicated by a discontinuity in regression lines at the cutoff. But in this case, because the regression lines are flat, this amounts to a test of the difference between group means on the postprogram knowledge measure. Here, the assignment measure is random with respect to the posttest and thus, by using uncorrelated pre-post measures, a type of random assignment has been achieved. Of course, such unrelated measures typically would not be used, because the need for a program is usually defined in terms of characteristics that might be affected by the program. However, there is nothing in the RD design

Figure 8. RD designs with uncorrelated pre- and postprogram measures



that requires identical or equivalent pre-post measurement.

This article has implied that the effects of a program can be examined on any number of outcome variables; two brief comments are in order. First, for each postprogram measure there is, in effect, a separate RD design. Each would be handled in a separate statistical analysis, and it cannot be assumed that a model that fits one postprogram measure will automatically be appropriate for another. Second, often it will be useful to analyze the total aggregate score for a given measure and then break the test into appropriately defined subtests for separate additional analyses. For example, if a particular health knowledge test provides a total score indicating the overall level of knowledge and that same test also can be divided into subtests that describe knowledge in separate areas (e.g., nutrition, diagnosis, and first aid), separate analyses could be conducted for both the total score and each subtest. In this way, the judicious selection of measures is likely to increase the probative value of an RD design and enhance the researcher's ability to suggest more specific program revisions.

In some contexts, postprogram measures will not be continuous normal variables. This complicates the modeling task, but does not necessarily preclude the use of RD designs. An excellent example is given by Berk and Rauma (1983), who used an RD design to evaluate the effects of a California law that extended unemployment benefits to released prisoners who previously had not been eligible. To qualify, prisoners had to earn at least $1,500 working in the prison during the final twelve months prior to their release. The State legislature was interested in whether extending such benefits would ultimately reduce recidivism in the form of parole revocation that usually resulted in a return to prison. Thus the key outcome measure was a dichotomous one where a score of 1 indicated recidivism and a 0 indicated none. Traditional regression analysis was clearly inappropriate, and visual inspection of the data for discontinuities at the cutoff would have been difficult. Berk and Rauma (1983) relied instead on a linear random utility model related to the binary logit model. These investigators concluded that ex-offenders who were in the program were about 10 percent less likely to return to prison. This example is instructive for health evaluation contexts. Dichotomous postprogram measures might be used to indicate mortality, recidivism, absence or presence of symptoms, and so on. While the use of postprogram measures that are not continuous normal variables is feasible in RD designs, the analyst must be cautious in constructing a statistical model to appropriately account for distributional form.

Statistical Analysis of the RD Design

The ability of an RD design to yield unbiased estimates of program effects depends on the degree to which the design has been implemented correctly and how well the true relationship between the assignment variable(s) and postprogram measure(s) has been modeled. A detailed description of implementation issues can be found in Trochim (1984). Discussions of analytic models are in Trochim (1984) and Judd and Kenny (1981). This section provides an overview of the assumptions behind a general analytic model, with some consideration of how this model might be altered for several of the design variations discussed earlier.

Assumptions of an RD analysis. Before a discussion of the specific analytic model is presented, it is important to understand the assumptions that must be met. The basic RD design as described earlier is assumed; variations in the design will be discussed later. There are five central assumptions that must be made in order for the analytic model presented here to be appropriate:

1. *The cutoff criterion.* The cutoff criterion must be followed without exception. When there is misassignment relative to the cutoff value (unless it is known to be random), a selection threat arises, and estimates of the program effect are likely to be biased. Misassignment relative to the cutoff, often termed a "fuzzy" RD design, introduces analytic complexities that are discussed in Trochim (1984) and Trochim and Spiegelman (1980).

2. *The pre-post distribution.* It is assumed that the pre-post distribution can be described as a polynomial function. If the true pre-post relationship is

logarithmic, exponential, or some other function, the model given below is misspecified and estimates of the program effect are likely to be biased. Of course, if the data can be transformed to create a polynomial distribution prior to analysis, the model may be appropriate, although it is likely to be more problematic to interpret. Sometimes, even if the true relationship is not polynomial, a sufficiently high-order polynomial will adequately account for whatever function exists. However, the analyst is not likely to know whether or not this is the case.

3. *Comparison group pretest variance.* There must be a sufficient number of pretest values in the comparison group to enable adequate estimation of the true relationship (i.e., pre-post regression line) for the group. It is usually desirable to have variability in the program group as well, although this is not strictly required because the comparison group line can be projected to a single point for the program group as discussed in Trochim (1984).

4. *Continuous pretest distribution.* Both groups must come from a single continuous pretest distribution, with the division between groups determined by the cutoff. In some cases it might be possible to find intact groups (e.g., two groups of patients from two different geographic locations) that serendipitously divide on some measure so as to imply some cutoff. However, such naturally discontinuous groups must be used with caution; because they differed naturally at the cutoff prior to the program, there is a greater likelihood that such a difference could reflect a selection bias that could introduce natural pre-post discontinuities at that point.

5. *Program implementation.* It is assumed that the program is uniformly delivered to all recipients (i.e., they all receive the same dosage, length of stay, amount of training, or whatever). If this is not the case, it is necessary to model explicitly the program as implemented, thus complicating the analysis somewhat.

A model for the basic RD design. The model presented here for the basic RD design is discussed in Trochim (1984) and is similar to the approach recommended in Judd and Kenny (1981). Given a pretest assignment measure, x_i, and a postprogram measure, y_i, the model can be stated as follows:

$$Y_i = b_0 + b_1 \tilde{x}_i + b_2 z_i + b3 \tilde{x}_i \tilde{z}_i + \ldots + b_{n-1} \tilde{x}_i^{\,s} z_i + e_i$$

where

$\tilde{x}_i$ = preprogram measure for individual i minus the value of the cutoff, x_o (i.e., $x_i = x_i - x_o$)

y_i = postprogram measure for individual i

z_i = assignment variable (1 if program participant; 0 if comparison participant)

s = the degree of the polynomial for the associated $\tilde{x}_i$

b_o = parameter for comparison group intercept at cutoff

b_1 = linear slope parameter

b_2 = program effect estimate

b_n = parameter for the s^{th} polynomial or interaction terms if paired with z

e_i = random error

The major hypothesis of interest is:

$$H_0: b_2 = 0$$

tested against the alternative:

$$H_1: b_2 \neq 0$$

There are several points to note about this model. First, the model estimates both main and interaction effects at the cutoff point–that is, the model looks for discontinuities in the pre-post polynomial relationship at the cutoff point. Second, the model requires that the analyst subtract the cutoff score value from each pretest score. The term $\tilde{x}_i$ has a superscript tilde to indicate this transformation of the pretest x_i. Finally, the model allows for any order of polynomial function (although certain restrictions are made in specifying the function, as described below). Thus, in theory, the true pre-post relationship can be linear, quadratic, cubic, quartic, and so on, or any combination of these.

Model specification. Given that the assumptions described above are correct and the general model is appropriate, the key problem in the analysis of an RD design is the correct specification of the polynomial model for the data at hand. Unfortunately, there is no simple or mechanical way to determine definitively the appropriate model for the data. Consequently, as with any statistical modeling, the RD analysis requires considerable judgment and discretion, since for any single RD design, the analyst will have to conduct multiple analyses based on different assumptions about the nature of the true pre-post relationship. The procedures outlined here explicitly encourage a multiple analysis approach.

Steps in the analysis. The following steps are necessary in specifying the model for an RD analysis. From a basic RD design, the analyst has a pretest assignment value (x_i in the model) for each person or unit. With this pretest score and the cutoff value, it is possible to create a new variable, z_i, which is equal to 1 for each subject who is in the program or 0 for those not in the program. This is accomplished on the computer with a simple transformation statement that recodes the pretest as a dummy-coded group membership variable in terms of the cutoff value. Finally, for each person there is a posttest score (labeled y_i in the model). Thus to begin the analysis there are three variables: x_i, z_i, and y_i. Given these variables, the steps used are as follows.

Transform the pretest. The analysis begins by subtracting the cutoff value from each pretest score, thus creating the term $\tilde{x}_i$ as in the model. This is done in order to set the intercept equal to the cutoff so that estimates of effect are made at the cutoff value (rather than at $x_i = 0$).

Examine the relationship visually. There are two major things to look for in a graph of the pre-post relationship. First it is important to determine whether there is any visually discernible discontinuity in the relationship at the cutoff. The discontinuity could be a change in level vertically (main effect), a change in slope (interaction effect), or both. If it is clear that there is a discontinuity at the cutoff, then analytic results that indicate no program effect should not be accepted as reliable. However, if no discontinuity is apparent, it may be that variability in the data is masking an effect, and the analytic results must be carefully examined. Secondly, the analyst should consider the degree of polynomial that may be required as indicated by the bivariate slope of the distribution, particularly in the comparison group. A good approach is to count the number of flexion points (i.e., number of times the distribution "flexes" or "bends") that are apparent in the distribution. If the distribution appears linear, there are no flexion points. A single flexion point could be indicative of a second- (quadratic) order polynomial. This information will be used to determine the initial model that will be specified.

Create higher-order terms and interactions. Depending on the number of flexion points detected, transformations of the transformed assignment variable, $\tilde{x}_i$, are created. The rule of thumb here is to go two orders of polynomial higher than was indicated by the number of flexion points. Thus if the bivariate relationship appeared linear (i.e., there were no flexion points), transformations up to a second-order (0 + 2) polynomial should be created. Since the first-order polynomial already exists in the model ($\tilde{x}_i$), only the second-order polynomial would have to be created by squaring $\tilde{x}_i$ to obtain $\tilde{x}_i^2$. For each transformation of $\tilde{x}_i$, the interaction term also is created by multiplying the polynomial by z_i. In this example there would be two interaction terms: $\tilde{x}_i z_i$ and $\tilde{x}_i^2 z_i$. Each transformation can be accomplished easily through straightforward multiplication on the computer. If there appeared to be two flexion points in the bivariate distribution, transformations up to the fourth (2 + 2) power and their interactions would be created. Visual inspection need not be the only basis for the initial determination of the degree of polynomial needed. Certainly, prior experience modeling similar data should be taken into account. The rule of thumb given here implies that the analyst should err on the side of overestimating the true polynomial function needed (for reasons outlined in Trochim, 1984). Based on the power initially estimated from visual inspection, all transformations and their interactions up to that power should be constructed. Thus, if the fourth power is chosen, all four terms $\tilde{x}_i$ to $\tilde{x}_i^4$ and their interactions should be constructed.

Estimate the initial model. At this point, the analysis can begin. Any acceptable multiple regression program can be used to accomplish this on the computer. The analyst simply regresses the posttest scores, y_i, on $\tilde{x}_i$, z_i, and all higher-order transformations and interactions created in Step 3. The regression coefficient associated with the z_i term (i.e., the group membership variable) is the estimate of the main effect of the program. If there is vertical discontinuity at the cutoff, it will be estimated by this coefficient. The significance of the coefficient (or any other) can be tested by constructing a standard t-test using the standard error of the coefficient that is invariably supplied in the computer program output. If the polynomial function required to model the distribution was correctly overestimated at Step 3, then the estimate of the program effect will at least be unbiased. However, by including terms that may not be needed in the true model, the estimate is likely to be inefficient–that is, standard error terms will be inflated; hence the significance of the program effect may be underestimated. Nevertheless, if the coefficient is highly significant at this point in the analysis, it would be reasonable to conclude that there is a program effect. The direction of the effect is interpreted based on the sign of the coefficient and the direction of scale of the posttest. Interaction effects also can be examined (e.g., a linear interaction would be implied by a significant regression coefficient for the $\tilde{x}_i z_i$ term).

Refine the model. The procedure described thus far is conservative with regard to bias. It is designed to reduce the chances of a biased program effect estimate even at the risk of increasing the error of the estimate. A full justification for this, which is outside the scope of this article, is provided in Trochim (1984). In brief, both an unbiased and efficient estimate is obtained if the specific model includes only the polynomial and interaction terms of the true relationship. Obviously, the analyst is never likely to know this true relationship with certainty.

If any term in the true relationship is omitted from the analysis (regardless of any other terms that are included), the model is considered underspecified and a biased estimate of program effect is likely. On the other hand, if all of the necessary terms from the true model are included in the analysis along with other unneeded ones, the model is considered overspecified. In theory, the unnecessary terms would be expected to have nonsignificant coefficients. In this case, the estimate of the program effect is unbiased, but the analyst "pays for" the inclusion of unnecessary terms with lower efficiency of the estimate. In general, the results of the initial model are likely to be overspecified and, while unbiased, may be inefficient.

On the basis of the results of Step 4, an attempt may be made to remove apparently unnecessary terms and reestimate the treatment effect with greater efficiency. This is a tricky procedure and should be approached cautiously to minimize the possibility of bias. To accomplish this, the output of the regression analysis in Step 4 should be examined, and the analyst should note the degree to which the overall model fits the data, the presence of any insignificant coefficients, and the pattern of residuals. A conservative method for deciding how to refine the model is to begin by examining the highest-order term in the model and its interaction. If both coefficients are nonsignificant, and the goodness-of-fit measures and pattern of residuals indicate a good fit, the analyst might drop these two terms and reestimate the resulting model. Thus, if a fourth-order polynomial was estimated and the coefficients for $\tilde{x}_i^4$ and $\tilde{x}_i^4 z_i$ were found to be nonsignificant, these terms could be dropped and the third-order model respecified. This procedure would be repeated until (a) either of the coefficients is significant, (b) the goodness-of-fit measure drops appreciably, or (c) the pattern of residuals indicates a poorly fitting model. The final model still may include unnecessary terms, but there are likely to be fewer of these and, consequently, efficiency should be greater. Model specification procedures that involve dropping any term at any stage of the analysis are more dangerous and more likely to yield biased estimates because of the considerable multicolinearity that will exist between the terms in the model.

Present the results. Because of the difficulties associated with model specifications in the RD design, usually it will be necessary to conduct multiple analyses of the same data under different model assumptions. While such a multiple analysis approach is becoming more of a standard among social science researchers (Reichardt and Gollob, 1986), it clearly runs counter to the typical administrator's desire to obtain the answer for a question. Apparently, there is no simple way out of such a dilemma. The analyst may choose to list the different estimates of the program effect by model, indicate the range of the estimates by reporting the highest and lowest values, or highlight a single estimate as a "best" or "favorite" one while also reporting the others.

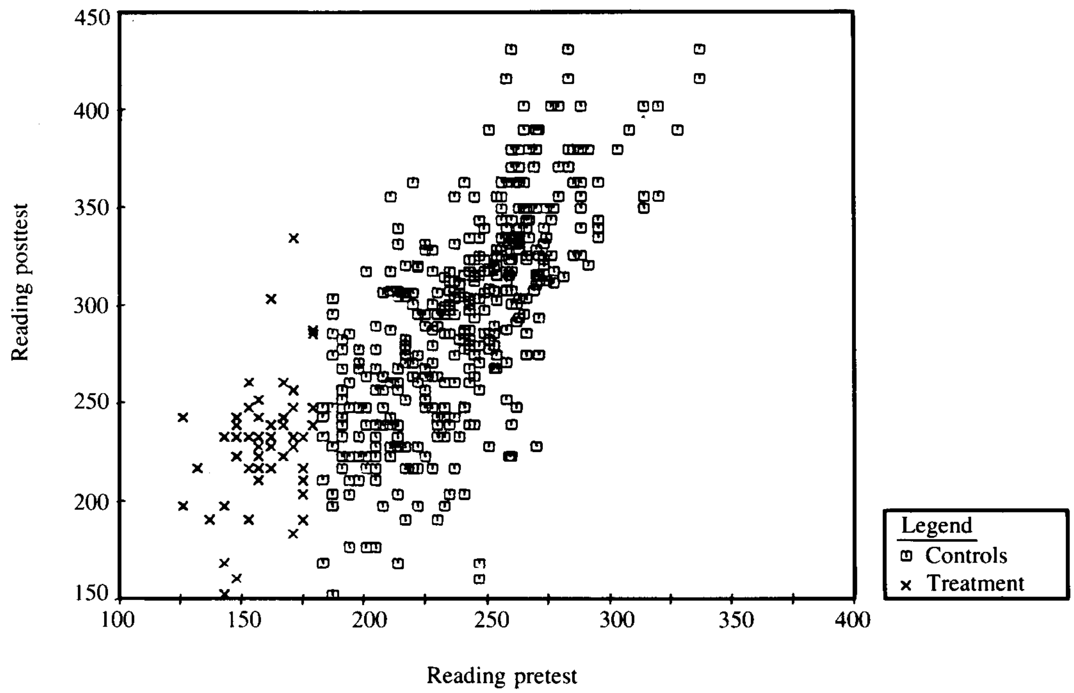
An illustrative analysis. To illustrate the steps in an RD analysis, data are presented from an evaluation of a compensatory education reading program. The program was conducted during the 1978-79 school year in the Providence, Rhode Island school district and was administered through the federally funded Title I of the Elementary and Secondary Education Act of 1965. The analysis presented here includes all second-grade students who were pre- and posttested and enrolled at an eligible school. The Comprehensive Test of Basic Skills (CTBS) was administered as the pretest in March 1987 and as the posttest in March 1979. This analysis and other examples are discussed in Trochim (1984), along with detailed consideration of some of the complexities of the Title I compensatory education evaluation context. As in the preceding discussion, the analysis comprised the steps described below.

Transform the pretest. For this program, the cutoff value was a CTBS pretest score of 179. Therefore, the first step in the analysis involved subtracting this value from each pretest value, $\tilde{x}_i$. The cutoff value divides this second-grade sample into two groups: the low-scoring program group, consisting of 54 cases, and the high-scoring comparison group, which had 411 students.

Examine the relationship visually. The bivariate distribution is shown in Figure 9. One of the first things that will be noticed is a fair amount of noise or variability in the data and some visual suggestion of outliers or "stray" points. Despite this, it should be visually apparent that there is a jump or discontinuity at the cutoff value. To see this, imagine some regression line that fits the comparison group cases. Now, extend that line to the left into the area of the program group. It should be seen that most of the program group cases have higher posttest values than that line would predict (i.e., fall above test values than that line would predict (i.e., fall above the line). On the basis of this rough visual assessment, the analysis would be expected to provide evidence for a program effect. Because of the direction of the discontinuity, it also would be expected that the program had a positive effect on the participants by increasing their posttest scores relative to what could be expected.

The second factor involves the degree of polynomial that might be appropriate for the model. Using the "flexion point" rule of thumb described earlier, it might be argued that there is a single flex point in the comparison

Figure 9. Example analysis: Bivariate distribution - Providence second grade reading



group distribution in the vicinity of a pretest score of 250 or so, suggesting that a quadratic term might be needed. There is little evidence for any higher-order model. It is worth noting that the program group has a relatively restricted pretest range that makes it difficult to assess visually what the slope of the program group line might be. In this case, there is some suggestion that the linear slopes of the two groups are different (a linear interaction). This will be examined in the statistical analysis.

Create higher-order terms and interactions. Following from the rule of thumb described earlier, this evaluation will begin with an analysis that goes two orders of polynomial higher than indicated by the number of flexion points. Since there was possibly one flex point, this means it will begin with an analysis that includes up to third-order (i.e., cubic) terms and their interactions. Specifically, the first model to be specified is as follows:

$$Y_i = b_o + b_1 \tilde{x}_i + b_2 z_i + b_3 \tilde{x}_i z_i + b_4 \tilde{x}_i^2 + b_5 \tilde{x}_i^2 z_i + b_6 \tilde{x}_i + b_7 \tilde{x}_i^3 z_i + e_i$$

Estimate the initial model. The parameter estimates for the initial model are shown in Table 1. The only term in this model that is significant is b_o, the constant. The program effect is estimated to be 21.69 CTBS points, with a standard error of 17.25. Thus the program effect is in the positive direction (as expected from examination of the bivariate distribution) but is not statistically significant. However, as described above and in Trochim (1984), the inclusion of terms that are not needed in the model lowers the efficiency of the program effect estimate (i.e., inflates standard errors and consequently underestimates significance).

Refine the model. A conservative way to revise the initial model would be to examine the highest-order term in the model and its corresponding interaction term to see if they can be eliminated. Here, this would involve removing the cubic and cubic interaction terms from the initial model and reestimating. These results are shown in Table 2 under the heading "Revision 1: cubic terms eliminated."

Here, the program effect estimate is again positive, but this time it is significant at a .05 level ($b_2 = 37.45$, SE[b_2] = 13.55, p = .006). However, as before, the highest-order terms in this model are not significant, and so it might be revised again, this time eliminating the quadratic terms. The results of this model are shown in Table

Table 1. Estimates for initial model, second-grade reading program, Providence, RI school district, 1978-79

Variable	b	SE(b)	p
Constant	216.27346	9.02495	<.001
Linear ($\tilde{x}_i$)	.88212	.49993	.078
Program effect (Z_i)	21.69478	17.24679	.209
Linear interaction ($\tilde{x}_i z_i$)	-2.52099	2.85144	.377
Quadratic ($\tilde{x}_i^{\ 2}$)	.00816	.00774	.292
Quadratic interaction ($\tilde{x}_i^{\ 2} z_i$)	-.13967	.13471	.300
Cubic ($\tilde{x}_i$ 3)	-.00003	.00003	.266
Cubic interaction ($\tilde{x}_i^{\ 3} z_i$)	-.00171	.00173	.324
	R^2 = .56919		

2, under the heading "Revision 2: cubic and quadratic terms eliminated." Here, little change is found in the program effect estimate. It is worth noting that the linear interaction term is not significant (although $p = .19$), and therefore there is no direct evidence for a difference in slopes between groups. Consequently, one more model revision might be tried, excluding the linear interaction term from the model. The results of this model are presented in Table 2, under the heading "Revision 3: linear term only." As before, this analysis provides evidence for a significant positive effect of the program on reading scores ($b_2 = 44.6$, $SE[b_2] = 7.5$, $p < .001$).

There are many other ways to examine the various models that might be specified. For instance, plots of predicted values or residuals would be especially useful when comparing models and trying to decide on revisions. Often it is useful to examine R-squares and analyses of various statistics. In this analysis, for instance, the R^2 for the initial model is almost identical to the R^2 for the final revision, indicating that little information was lost when the terms were eliminated during revision.

In summary, this illustrative analysis began by noting the visual evidence of a discontinuity indicative of a positive program effect. Also, some slight evidence of nonlinearity in the distribution was noted that led initially to specification of a third-order RD model. The program effect estimate for this model was indeed positive but not statistically significant. In all subsequent revised models where attempts were made to eliminate likely unneeded terms, the program effect was positive and statistically significant, and there was little evidence of loss of information. On this basis, it would be reasonable to conclude that the analysis supports the hypothesis that the program had a positive effect on reading scores.

Some comments on design variations. The analytic model reported above is stated in terms of the basic RD design. However, it can be revised easily and extended to other RD variations discussed in this article. Each of the major variations presented earlier will be discussed briefly.

Analyses with multiple cutoff points. When multiple cutoffs are used, multiple analyses will be necessary. For the simplest case of two cutoff values (as in Fig. 6), three types of analyses will be useful. First, all cases within the cutoff interval (i.e., between the cutoffs) can be eliminated, and the program effect can be estimated. Here, the effect of the program probably would be estimated at the cutoff value that separates the RD comparison group from the rest of the cases (i.e., at a score = 4 in Fig. 6). Second, all cases falling outside of the cutoff interval can be excluded, and the appropriate analysis can be conducted on the remaining cases. Thus, if there is random assignment within the interval, the analyst could conduct a traditional ANCOVA analysis. If the within-interval groups are nonequivalent, appropriate statistical adjustments for selection bias might be attempted. Finally, all cases would be included in a single analysis. Again, it is probably most sensible to estimate the effect of the program at the cutoff point that places all program participants on one side. Thus, in Figure 6, it makes more sense to estimate the program effect at $x = 4$ than at $x = 5$, because some program participants fall within the interval of 4-5, whereas no program cases have a pretest less than 4. Thus in both RD analyses of the data above, the analyst typically would subtract the comparison group cutoff value rather than the program group cutoff because there might be an interaction between pretest and program that would change the shape of the regression function. Since any such interaction presumably would affect all program participants, it is important that the estimate be calculated at a point on the pretest where this interaction is likely to commence.

In any case where there are multiple cutoffs and more than two groups (as in Fig. 7), it will be desirable to estimate many different program comparisons. For instance, in the example shown in Figure 7, the analyst might combine all program cases at each wave of the program and do a basic RD analysis, or the analyst could keep different program groups separate in order to examine long-term effects of the program. In this latter case, multiple program assignment variables and all necessary interaction terms would have to be constructed for the analytic model (e.g., z_1, z_2, z_3). Clearly,

Table 2. Estimates for revised models, second-grade reading program, Providence, RI school district, 1978-79

Variable	b	SE(b)	p
Revision 1: cubic terms eliminated			
Constant	209.1262	6.3448	<.001
Linear ($x_{\tilde{i}}$)	1.3919	.2006	<.001
Program effect (Z_i)	37.4535	13.5511	.006
Linear interaction ($x_{\tilde{i}} z_i$)	-.4602	1.2013	.702
Quadratic ($x_{\tilde{i}}^2$)	-.0003	.0014	.836
Quadratic interaction ($x_{\tilde{i}}^2 z_i$)	-.0023	.0246	.925
	R^2 = .56706		
Revision 2: cubic and quadratic terms eliminated			
Constant	210.11	4.22	<.001
Linear ($x_{\tilde{i}}$)	1.35	.06	<.001
Program effect (Z_i)	35.84	10.06	<.001
Linear interaction ($x_{\tilde{i}} z_i$)	-.51	.39	.193
	R^2 = .56702		
Revision 3: linear term only			
Constant	210.87	4.18	<.001
Linear ($x_{\tilde{i}}$)	1.34	.06	<.001
Program effect (Z_i)	44.61	7.50	<.001
	R^2 = .56542		

as more cutoffs and groups are added, the model-specification issues are complicated.

Analyses with multiple assignment measures. Two versions of multiple assignment measure RD designs were discussed. The first has multiple measures, each of which has its own cutoff. In this case each assignment measure would be transformed by having its own cutoff value subtracted from it to create $x_{\tilde{i}}$. In the analysis, all transformed assignment measures, group membership, higher-order terms, and interactions would be included. For the simple first-order (or linear case) with these assignment variables, the analyst could use the model:

$$Y_{\tilde{i}} = b_o + b_1 z_i + b_2 x_{\tilde{i}} + b_3 x_{2\tilde{i}} + b_4 x_{1\tilde{i}} z_i + b_5 x_{1\tilde{i}} z_i + b_6 x_{2\tilde{i}} z_i + b_7 x_{3\tilde{i}} z_i + e_i$$

This model does not include any two-way assignment variable interaction terms (e.g., $bx_{1\tilde{i}}x_{2\tilde{i}}$ or $bx_{1\tilde{i}}x_{2\tilde{i}} z_i$) or any three-way terms (e.g., $bx_{1\tilde{i}}x_{2\tilde{i}}x_{3\tilde{i}}$ or $bx_{1\tilde{i}}x_{2\tilde{i}}x_{3\tilde{i}}z_i$), but such an assumption may not be reasonable because the primary interest is in the estimate of b_1, the program effect, not in interactions per se. Nevertheless, it should be apparent that the use of multiple assignment variables with higher-order polynomial models will quickly lead to an unwieldy analysis.

The situation is simpler when multiple assignment variables and an "or" assignment rule are used. In this case the real assignment variable might actually be labeled the "number of assignment variables on which a person meets the criterion." Thus, with 10 assignment variables, there is a single preprogram measure with values between 0 and 10, where the value for any person indicates how many of the 10 separate assignment criteria are met. With this as the x_i variable, the basic RD analysis as described earlier would be conducted. When multiple assignment variables are combined into a single index and a cutoff on that index is assigned, the analysis is also a straightforward application of the basic analysis described initially.

Conclusion

The following, simple guidelines could be used to make a preliminary assessment of the appropriateness or feasibility of an RD design for a given study. As with all simplifications, there are potential hazards here, and these guidelines should not be considered a substitute for a more careful reading of this article and the other sources cited. Three basic guidelines are suggested.

1. **Program guideline.** The purpose of the research must be to assess the effect of some program or treatment. As simple as this may sound, often it is not a trivial consideration. The purpose of the study must be to investigate a causal hypothesis; do not let the research design define the research question. Just because a program is assigned by a cutoff strategy doesn't mean there is interest in evaluating that program. The researcher must be especially careful that the cutoff-based assignment is used for the program that is to be evaluated. For instance, if patients are admitted to a special hospital unit on the basis of a cutoff on a severity of illness indicator, it may be possible to use the RD design to evaluate the effect of that entire unit on subsequent illness. However, if the interest is in assessing the effects of a new surgical technique that is given only to some of the people admitted to the unit (and solely on the basis of physician discretion), then it doesn't really help that unit assignment is cutoff based. It is important to be clear about what program or programs are being evaluated and their relationships to the cutoff rule.

2. **Assignment guideline.** Persons (or units) must be assigned to the program solely on the basis of a cutoff score; this is the distinguishing characteristic of the RD design. There must be no exceptions to this rule, or the quality of the RD design will be jeopardized. When trying to assess the feasibility of the RD design, the analyst should not give up on it simply because a single cutoff value on a single preprogram indicator is not feasible. There can be many different variations that might work in a given situation–multiple cutoffs on a single assignment measure, multiple assignment measures with their own cutoffs, aggregate indices, and so on–but assignment to the program must be cutoff-based.

3. **Measurement guideline.** There must be (a) measurement of all cases–program and comparison–on both pre- and postprogram measures and (b) sufficient data to estimate a regression line reasonably, at least for the comparison group. Often, the first requirement is extremely difficult to achieve in practice. Frequently, those who are denied the program (the comparison group) are not tracked after the program is given, and there is consequently no postprogram measurement on them. Measuring the comparison group can be an expensive proposition. In some instances, it may be reasonable to sample randomly from the comparison cases rather than to measure the entire group, as described by Trochim (1984). Secondly, there must be sufficient variability on the preprogram measure to enable the estimation of a regression line. Even if this is not possible for the program group (as was nearly the case in the simulation of the reading program described earlier, where there was almost too little program group pretest variability to warrant fitting a line through that group), it must be the case, at least minimally, for the comparison group.

When considering the possibility of an RD design, each of the guidelines should be examined to ensure that the necessary minimal program, assignment, and measurement conditions are met. If this appears to be the case, it would then be reasonable to examine the circumstances in greater detail in order to set up the design correctly.

RD designs appear to have great promise for evaluation in health contexts. They stand as perhaps the strongest alternative to randomized experiments. Given the wealth of quantitative indicators and the tendency of many medical and health-related professionals to use such indicators as the basis for making decisions, RD designs may be directly applicable or usable with only minor modifications to existing procedures.

However, RD designs are not a panacea. They are based on some fairly restrictive assumptions: that assignment can be based solely on cutoffs, that more than just the recipients of the program can be measured, that there is enough premeasure variation in the comparison group to estimate the true pre-post relationship, and so on. In many contexts, one or more necessary conditions are likely to be absent, thus ruling out the use of this design. But when all conditions are met, the RD design should be considered a strong method for evaluating health programs.

References

American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). Washington, DC: Author.

Berk, R.A. and D. Rauma. (1983). Capitalizing on nonrandom assignment to treatment: A regression discontinuity of a crime program. *Journal of the American Statistical Association*, 78, 21-28.

Boruch, R.F. (1975). Coupling randomized experiments and approximations to experiments in social program evaluation. *Sociological Methods and Research*, 4, 31-53.

Campbell, D.T. (1986). Relabeling internal and external validity for applied social scientists. In W.M.K. Trochim (Ed.), Advances in quasi-experimental design and analysis [Special issue]. *New Directions for Program Evaluation*, 31, 67-77.

Campbell, D.T. and J.C. Stanley. (1963). Experimental and quasi-experimental designs for research on teaching. In N.L. Gage (Ed.), *Handbook of research on teaching*. Chicago: Rand McNally.

Campbell, D.T. and J.C. Stanley. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Cook, T.D. and D.T. Campbell. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.

Draper, N.R. and H. Smith. (1981). *Applied regression analysis*. New York: John Wiley and Sons.

Goldberger, A.S. (1972). *Selection bias in evaluating treatment effects: Some formal illustrations* (Discussion Papers, 123-72). Madison: University of Wisconsin, Institute for Research on Poverty.

Judd, C.M. and D.A. Kenny. (1981). *Estimating the effects of social interventions*. New York: Cambridge University Press.

Reichardt, C.S. and H.F. Gollob. (1986). Satisfying the constraints of causal modeling. In W.M.K. Trochim (Ed.), Advances in quasi-experimental design and analysis [Special issue]. *New Directions for Program Evaluation*, 31, 91-107.

Tallmadge, G.K. and C.T. Wood. (1978). *User's guide: ESEA Title I evaluating and reporting system*. Mountain View, CA: RMC Research Corporation.

Trochim, W.M.K. (1984). *Research design for program evaluation: The regression-discontinuity approach*. Beverly Hills, CA: Sage.

Trochim, W.M.K. (1985). Pattern matching, validity, and conceptualization in program evaluation. *Evaluation Review*, 9, 575-604.

Trochim, W.M.K. and C.H. Spiegelman. (1980). The relative assignment variable approach to selection bias in pretest-posttest group designs. *Proceedings of the Social Statistics Section, American Statistical Association.*

Comment

The Applicability of the Regression-Discontinuity Design in Health Services Research

Harold S. Luft, Ph.D.

The description of the regression-discontinuity design provided by William Trochim (in this volume) is quite exciting. It offers a valuable addition to the standard randomized controlled trial (RCT) as a method of allocating subjects to treatment and nontreatment groups. In fact, in some ways it may be preferred to the RCT which is currently the "gold standard" in medical research. It is clear that significant health policy results often carry greater weight when derived from RCTs than from other methods. This is evidenced by the obvious weight given the findings from the Rand health insurance study (Newhouse, Manning, Morris, and others, 1981) with its randomized design.

It should be pointed out that the regression-discontinuity (RD) design is applicable only when there is controlled assignment of subjects to treatment and nontreatment groups. It is not appropriate for the analysis of simple observational data, nor is it appropriate for situations in which there is uncontrolled nonexperimenter discretion in assignment. However, in appropriate circumstances, RD appears potentially valuable.

Two major advantages of the RD design are highlighted here. The first is its ability to address certain ethical problems that may arise in the application of the classic RCT design. The potential ethical advantage is mentioned in Trochim's paper, but this discussion will explain how an economist also could apply the RD design. The second advantage is more subtle and perhaps more controversial, since it suggests a practical analytic advantage of the RD design over the RCT. It is clear that if an RCT can be undertaken, the study design will be stronger, and the results will be more readily accepted than with the RD approach. However, if an RCT is not feasible due to ethical or logistical reasons, RD may provide a useful alternative.

In the basic RD approach, subjects are assigned to treatment and nontreatment groups based on some cutoff score on a pretreatment measure. This assignment process has obvious advantages in situations such as the clinical trial of a new drug. If there are plausible arguments to believe the new drug is beneficial, it may be difficult ethically to randomly withhold it from some individuals merely to prove its efficacy. Consider for example the situation of AIDS patients in the trials of AZT. The RD design allows researchers to offer the drug to all those considered most severely ill, as long as there is some reasonable way of scoring severity.

Similar situations can arise in more classic health services research problems. For example, a recent project was undertaken to design an intervention that would provide feedback to hospitals concerning their quality of care based on the analysis of routinely available patient discharge abstracts or billing data. Originally, a randomized design was considered in which only half the hospitals identified as having potential quality of care problems would be offered the information and consultation to help interpret the data. While this classic RCT (albeit without blindness) would provide the clearest measures of the effectiveness of the feedback intervention, there was concern about the ethics of withholding data about potential quality of care problems from half of the hospitals.

In contrast, the following RD design provides a relatively simple solution to this ethical dilemma. The initial analysis of the data would produce a score indicating the probability that the number of poor outcomes occurring at each hospital would be observed by chance if the hospital truly had average quality of care, given its case mix. This is a natural pretreatment score, and feedback could be provided to all hospitals with scores above a certain level, such as a Z-score greater than 1.96. Alternatively, since there was concern also about the resources that would be required to provide the on-site consultation to explain the data, the offer of feedback could begin with those hospitals having the highest Z-scores and work down the list until consultation time was exhausted. This would not guarantee the provision of feedback to all hospitals that could possibly benefit, but it would provide at least an objective and ethical way of allocating limited resources.

Dr. Luft is Professor of Health Economics in the Institute for Health Policy Studies, University of California at San Francisco.

In other situations, administrative decisions may lend themselves to the application of an RD design. For example, some State Medicaid programs have begun to use case managers to control the medical care use of people with patterns of high utilization. Similar systems have been implemented by some insurers for certain employers. The often naive evaluations of such programs compare pre- and postuse for persons enrolled in the case management program. Such evaluations receive short shrift from the slightly more sophisticated analyst who recognizes that after choosing people because of their high use patterns, regression to the mean is likely to result in decline in use even if case management had no effect. Given the nature of the programs, it is politically and administratively infeasible to randomly assign high users to managed care and usual coverage groups. However, if data on people assigned to the program also are available, which is usually the case in such programs, then an RD design could be used to determine if the reduction in use is greater than would be expected due to regression to the mean.

One of the technical problems with RCTs arising from the ethical issues related to randomization is the need to focus attention on that group of individuals (or organizations) for whom the treatment is neither clearly beneficial nor clearly unnecessary, based on prior expectations. Thus, a well designed RCT often considers for randomization only those people in the mid-range of some pretreatment criteria, such as patients classified as mild hypertensives rather than normotensives or those with high blood pressure. This narrow focus may necessitate a longer period to accrue enough subjects or require complex multicenter collaborations. Furthermore, the narrow range of subjects makes it difficult to determine covariates that may enhance or reduce the treatment effect. The RD approach, in contrast, allows the inclusion of a much broader range of subjects, possibly counteracting the reduced power of the design compared with the RCT with a lower cost per subject. Much more work is needed to examine the costs of achieving equally credible results using alternative designs under various situations.

The second advantage of the RD design is actually the flip side of its major weakness, the need to adequately model the true relationship between the assignment variable and the posttreatment outcome measure. In a sense, the heart of the RD design is in identifying whether a discontinuity or break in the relationship occurs at the point differentiating the treatment and nontreatment groups. This discontinuity may be a shift in the intercept, a change in slope, or some other more complicated relationship. However, if the underlying relationship is not well understood, it may be difficult or impossible to determine whether the treatment altered the effect. This means that much more attention must be given to the underlying relationship between pretest and posttest measures than is the case in an RCT when just the posttest measures are necessary. Health services researchers, however, often spend much of their time developing such models using nonexperimental data. Applying such expertise to the RD approach in an experimental situation may have a substantial payoff. This is an example of the application of "little t-theory" to the design and analysis of the data.

For example, suppose that the treatment under study has an effect related to the assignment score, which implies a shift in slope for the treatment group rather than just a change in intercept. This is probably a fairly frequent phenomenon. In the quality of care study described above, hospitals with very high Z-scores may be more likely to have real quality of care problems. In the case of hospitals with lower scores, the likelihood that the observed outcomes were due to chance increases, so feedback would have little effect. By forcing the analyst to think about and focus on the underlying relationship, the RD design may be more likely to point to where interventions are more or less effective. It may even be easier to identify an effect with an RD than an RCT design. If a classic RCT had been undertaken, say by splitting the sample of eligible hospitals in half, it may have been more difficult to detect an effect because of the dilution of the treatment group by the inclusion of hospitals with relatively low Z-scores (for which there is little effect) and, conversely, by the inclusion of high Z-score hospitals in the control group.

The above comparison implies a fairly sophisticated RD design and analysis in contrast to a rather simple RCT design, which of course is unfair. Adjustors can be added to an RCT and similarly increase its power, but such an approach is sometimes difficult for readers and especially policymakers to understand and accept. A major advantage of the RCT is its simplicity; once the analysis has been complicated, it appears to some that the data are being "cooked" to obtain a desired result. At this point, the RD design may begin to appear even more straightforward and thereby overcome one of the obstacles to its use that is based on its purported complexity and lack of "face validity."

In conclusion, the regression-discontinuity design seems a worthy addition to the set of tools available to the health services researcher. In particular, it provides a way to avoid the ethical dilemma of withholding treatment to those most in need, and it addresses the regression to the mean problem in programs focusing on "outliers" or high users. The RD approach also relies on the careful wedding of analysis and design, which is likely to lead to improved research.

Reference

Newhouse, J.P., W. Manning, Jr., C. Morris, and others. (1981). Some interim results from a controlled trial of cost-sharing health insurance. *New England Journal of Medicine,* 305, 1501-1507.

Regression-Discontinuity Design in Health Evaluation

Sankey V. Williams, M.D.

Trochim (in this volume) claims that the regression-discontinuity (RD) design is "...perhaps the strongest alternative to randomized experiments. . . ." The implication is that the RD design should replace alternative nonrandomized designs in some studies and could replace randomized clinical trials in other studies. Trochim is right–but only partly right.

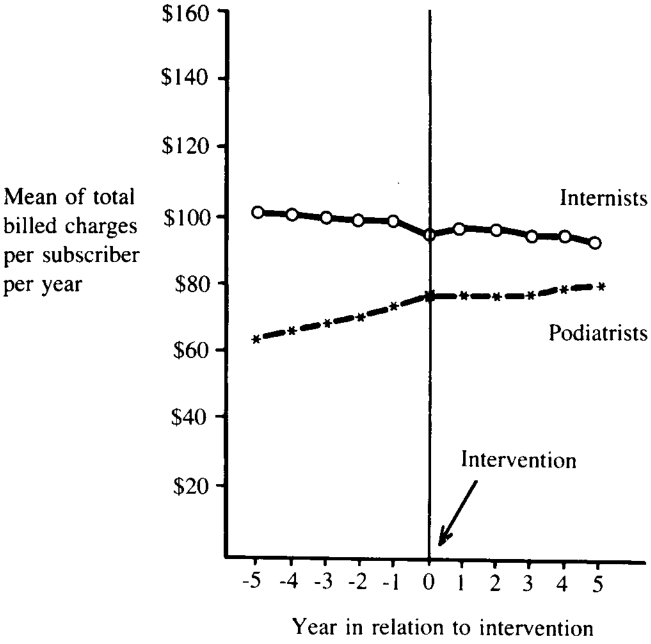
No doubt all researchers have projects that cannot be completed because of methodological problems, and two such projects will illustrate some of the advantages and disadvantages of the RD design.

The first project attempted to determine whether cost-control measures imposed by Pennsylvania Blue Shield were effective (Schwartz, Williams, Eisenberg, and Kitz, 1982). Pennsylvania Blue Shield is the largest Blue Shield organization in the country, and it is the fiscal intermediary for Medicare in Pennsylvania. Consequently, it pays bills from most of the State's physicians. To control costs, Pennsylvania Blue Shield developed a unique monitoring program. Each physician's total yearly charges were calculated separately for each service; for example, total charges were calculated separately for the physican's office visits, hospital consultations, and interpretations of electrocardiograms. Individual physicians then were combined into peer groups depending on their specialty. The distribution of charges in each peer group was examined to determine the cutoff value below which 95 percent of physicians could be found. This process identified the 5 percent of physicians who had the highest total charges for each service in each peer group. Finally, for every physician Pennsylvania Blue Shield calculated the total cost of all services that was above the 95 percent cutoff values. If the total was $5,000 or greater, the doctor received an informational letter that described the calculations, identified the physician as having unusually high charges, and promised no further action (which may or may not have been believed).

Because of problems with extracting data, only internists and podiatrists were included in the study. Of all internists studied, 405 did and 1,400 did not receive a letter. Of all podiatrists, 85 did and 709 did not receive a letter. The study was concerned with the effects of these letters on subsequent charges. If the letters were effective, subsequent charges should have decreased. The study had a pre-post design with a nonrandomized control group.

The control group consisted of physicians who did not receive a letter but did bill for services in the same time period as those who received a letter. In Figure 1, the horizontal axis indicates the number of years before and after receipt of the letter. The vertical axis refers to total charges (per patient), adjusted for inflation. In neither control group was there any confounding effect that could be related to receipt of the letter by matched physicians in the study group (in Fig. 1, the letter was received at time zero as indicated by the vertical line).

Figure 1. Pattern of charges for control groups



Dr. Williams is Director of the Section of General Internal Medicine, Leonard Davis Institute of Health Economics, and Professor of Medicine and Health Care Systems, University of Pennsylvania.

Figure 2. Pattern of charges for internists and podiatrists who received a letter

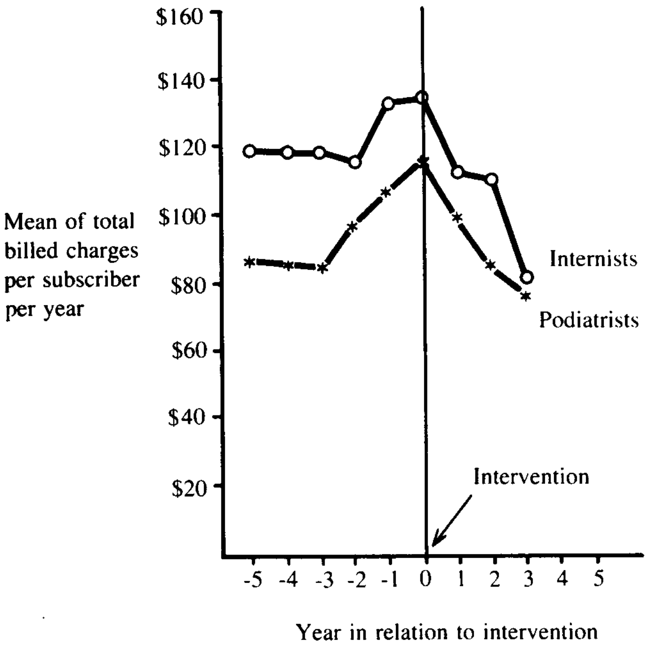
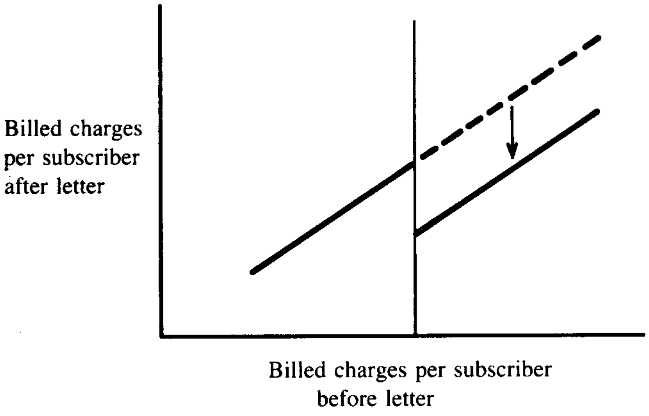


Figure 3. Expected pattern of charges for combined control and letter groups with the RD approach


Billed charges per subscriber after letter
Billed charges per subscriber before letter

In contrast, podiatrists and internists who eventually received letters increased their charges each year until they reached the cutoff value (Fig. 2). After receiving a letter, they decreased charges sharply, and the effect persisted for as long as they were followed. When regression lines are calculated separately for pre- and postletter values and the slopes are compared, the differences are statistically significant. However, the design has been criticized because it does not distinguish regression to the mean from the effect of the letter. Critics argue that those with the top 5 percent of charges before the letter are more likely to regress toward the mean than those in the control group, which might explain some or all of the letter's apparent effect.

The RD design may provide a solution. If the data were reanalyzed using the RD design, the results might be as diagrammed in Figure 3. Here, the horizontal axis describes billed charges before receipt of the letter, and the vertical axis describes billed charges after receipt of the letter. The graph includes data from all physicians–both those who did receive a letter and those matched for the same specialty and time period who did not receive a letter. The higher the billed charges before the letter, the higher they would be expected to be after the letter. If there had been no letter, the regression line would follow the dashed line. If the letter were effective, the discontinuity represented by the solid line to the right of the vertical line should be observed.

The beauty of this new analysis is that it is not biased by regression to the mean. According to Trochim, the low-scoring pretest group is expected to evidence a relative gain on the posttest and the high-scoring pretest group to show a relative loss. However, it is not expected that regression to the mean will result in a discontinuity in the bivariate relationship coincidental with the cutoff point. If Trochim is correct, the data can be reanalyzed to distinguish between regression to the mean and a letter effect.

In this case, Trochim was right. The RD design has superior methodological properties that make it more useful than the pre-post design with nonrandomized control groups.

The second project occurred several years ago. A randomized clinical trial was planned to determine if patients with heart attacks should be cared for in special coronary care units, which is standard practice, or whether they could be cared for on standard wards with telemetry devices that monitored for cardiac arrhythmias. Arrhythmias are the chief preventable cause of death in acute heart attacks, and this project was intended to show that arrhythmias could be detected and treated equally well in both places. The trial seemed important because coronary care units never had been shown to be effective in randomized clinical trials, new telemetry devices that allowed computer monitoring had been developed, there were theoretical reasons why special units might actually induce arrhythmias, and there were important cost differences between the two types of care. It was proposed that all heart attack patients would be admitted to the coronary care unit for the first 24 hours. During this period, an index would be calculated to predict the probability of hospital death. Patients with mild heart damage who fell below a cutoff value on the index would be randomized either to stay in the unit or to be transferred to ward telemetry. The outcome measure was hospital mortality.

A proposal for the trial was submitted to the National Heart, Lung, and Blood Institute. It was reviewed by the Clinical Trials Review Committee who liked it well enough to conduct a site visit. In the end, however, the proposal was not funded. The decision was close enough that the whole episode became the subject of a case study conducted by RAND and funded by the National Center for Health Services Research and Health Care Technology Assessment (NCHSR) on the problems of funding studies of standard medical practice (Hammons and Kahan, 1985).

There were a variety of problems. This proposal had all the problems that plague most other randomized clinical trials that look at important medical-practice issues. It was big. Because only small differences in death rates were expected in a population that had a low death rate to begin with, power calculations indicated that 2,300 patients were needed to have a reasonable chance of success. This many patients required that eight hospitals had to be involved, greatly increasing the project's complexity. Even with eight institutions, 4 years would be needed to complete the study, which meant the results would be slow in coming, perhaps slow enough that practice would change in the interim and make the question obsolete. To pay for enrolling so many patients over such a long time, over $5 million would be needed, which was expensive by NIH standards. Finally, there was the ethical issue; given what was and was not known, was it ethical to randomize patients? Perhaps surprisingly, this was not much of an issue in this instance. Most observers believed that about 50 percent of eligible patients would consent to be randomized.

According to Trochim, "... the [RD] design is a strong competitor to randomized designs when causal hypotheses are being investigated." What would have changed if, instead of randomizing, the RD design had been adopted? In the RD design, all patients with mild heart attacks would have been transferred to ward telemetry, and the results might have been those in Figure 4. Points on the graph are defined by the predictive-index value, which is on the horizontal axis, and the hospital death rate, which is on the vertical axis. Lower index values should be associated with lower death rates, regardless of where the patient is cared for. The dashed line describes what would have been observed if there were no differences between the special unit and ward telemetry, which is what was expected. The two solid lines to the left of the vertical line describe what might have been found if there were differences. If ward telemetry was better than the special unit, the lower line would have been observed; if it was worse, the upper line would have been observed.

It does not appear that the RD approach would have solved any of the problems associated with this study. Citing the work of others, Trochim states that, "... up to two-and-a-half times as many participants are needed in a [RD] design as in a randomized experiment in order to obtain comparable levels of statistical precision." With the RD design, more eligible patients probably could have been included, and in addition, information from patients with high scores could have been used. However, little would have changed. With the RD design, 4,000-5,000 patients would have been required, instead of the 2,300 patients needed for the randomized design. As a result, about the same number of institutions would have been required with all their complexity, and it would have taken about the same number of years to enroll all the patients. The price would not have changed very much. The ethical issues associated with randomization would have been avoided, but ethical issues did not prevent this proposal from being funded.

The RD design does not solve the problems posed by randomized clinical trials of medical practice, because its inefficiency fails to reduce the size, complexity, duration, or expense of these trials. The RD design might be an attractive alternative, however, when substantially fewer than 50 percent of eligible patients can be randomized. Only then would the ability to look at data from all patients compensate for the relative inefficiency of this design.

The RD design is not supposed to replace all randomized clinical trials. The design is especially promising when ethical considerations make it difficult to randomize patients, which was not an issue in the preceding example.

The following example is dominated by ethical issues. Consider a situation in which a promising new drug has been developed for the treatment of acquired immune deficiency syndrome, AIDS. The drug has been found to prevent replication of human immunodeficiency virus (HIV) in the laboratory. It has undergone short-term, uncontrolled trials in a small number of humans. These trials have defined the drug's absorption, volume of distribution, and metabolism and found few side effects. It is time for larger clinical trials that will test clinical effectiveness and measure the true incidence of side effects, but the drug is so promising and the need for effective therapy is so great that randomization is considered unethical by some.

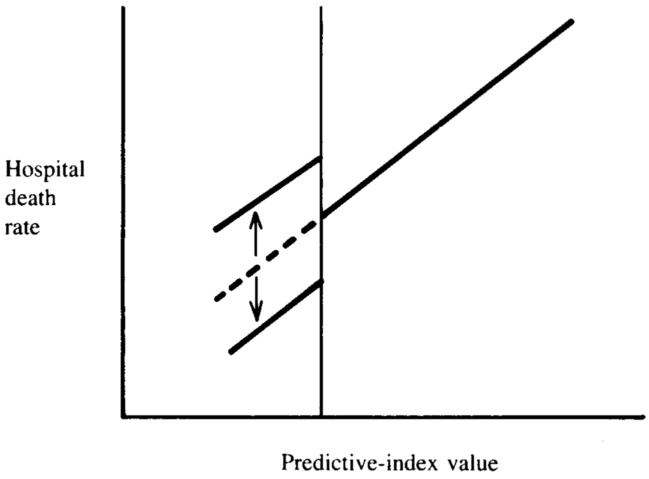
Would the RD design solve this problem? Candidate patients could be assigned an immunodeficiency score based on the number and proportion of specific lymphocytes in their circulating blood. A cutoff value could be determined to define the level of immunodeficiency that would qualify a patient for the new drug. All patients below this level would receive the new drug, and all those above the level would receive conventional therapy. The outcome variable could be the immunodeficiency score after a defined course of therapy, the number of opportunistic infections in a defined time period, or death.

How would patients and their physicians respond? Patients whose immunodeficiency score was just above the cutoff level would be denied the drug. These patients and their physicians would argue for exceptions to be made, and they would seek alternative sources for the drug, including black market sources if recent experience is a guide. Although the RD design may relieve some of the ethical pressure, it does not solve the problem. It merely shifts concern away from patients who are randomized to receive placebo to those who have immunodeficiency scores just above the cutoff value. If some of these patients were successful in getting the drug, either through an exception in the study or by finding an alternative source outside the study, the cardinal assumption of the RD design would be violated. The design would be converted into a "fuzzy" design; fuzzy designs cause problems because they produce biased results. Trochim has proposed the use of a "relative assignment variable" to deal with fuzzy designs (Trochim, 1984). The relative assignment variable has been shown to adjust for assignment bias in a small number of studies that used simulated data. However, too little is known about this novel approach for it to be used in a study where the results would be used to define life-or-death therapy for tens of thousands of potential patients.

Nevertheless, if the integrity of a randomized clinical trial could be maintained, it is possible that the integrity of an RD trial could be maintained as well. Pressure to create exceptions to the RD assignment rule probably would be no more intense than pressure to violate random assignment. Therefore, the RD design probably could be implemented without the need for a fuzzy analysis.

Given that a true RD trial could be implemented and analyzed without fuzzy methods, is it a feasible replacement for a randomized clinical trial?

Figure 4. Expected results of the myocardial infarction study with the RD approach



The RD design is less efficient than the randomized clinical trial. Thus, more patients will have to be included in an RD design than in a randomized clinical trial. If the drug is eventually found safe and effective, more patients will have been denied optimal care in an RD design than in a randomized clinical trial. If the drug is found to have unacceptable side effects for the level of effectiveness, more patients will have been exposed to the risk of side effects in an RD design than in a randomized clinical trial. Either way, more patients will be given the wrong therapy in an RD design than in a randomized clinical trial.

The RD design does not present a very attractive alternative to the randomized clinical trial, even when ethical issues dominate. The only exception is when ethical issues prevent a randomized trial from being performed at all.

In summary, it appears that Trochim is correct in asserting that the RD design can and perhaps should replace other nonrandomized designs in health services research. This is useful information that is not widely known. For example, a MEDLINE search revealed no studies–either in the medical literature or in the health services literature–that used the RD design. However, because of its relative inefficiency, the RD design cannot compete with the randomized clinical trial, except when fewer than 50 percent of eligible patients can be randomized, including the special case in which so few patients can be randomized that a randomized trial cannot be conducted.

References

Hammons G.T. and J.P. Kahan. (1985). *Initiating clinical trials: A case study of a proposed clinical trial for acute myocardial infarction,* [a RAND Note prepared for NCHSR]. Santa Monica, CA: RAND.

Schwartz, J.S., S.V. Williams, J.M. Eisenberg, and D. Kitz. (1982). Effect of utilization review (UR) on physicians' billings. *Clinical Research,* 30, 306A.

Trochim, W.M.K. (1984). *Research design for program evaluation.* Beverly Hills, CA: Sage.

AHCPR CONFERENCE PROCEEDINGS

RESEARCH METHODOLOGY: STRENGTHENING CAUSAL INTERPRETATIONS OF NONEXPERIMENTAL DATA

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
Public Health Service
Agency for Health Care Policy and Research

CONFERENCE PROCEEDINGS

RESEARCH METHODOLOGY: STRENGTHENING CAUSAL INTERPRETATIONS OF NONEXPERIMENTAL DATA

Edited by
Lee Sechrest, Ph.D
Edward Perrin, Ph.D
John Bunker, M.D.

This research was supported by AHCPR Grant No. HS 05306

May 1990

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
Public Health Service
Agency for Health Care Policy and Research

Contents

Comment

Feature Article

Comment

Feature Article

Comments

Feature Article

Comment

Feature Article

Comment

Summaries