Focus on Research Methods

Threats to Validity in Randomized Clinical Trials

Louis Fogg,1* Deborah Gross2†

1Department of Psychology and Psychiatry, Rush–Presbyterian–St. Luke’s Medical Center, Chicago, IL
2Rush University College of Nursing, 600 South Paulina, Suite 1042B, Chicago IL 60612
Received 22 December 1998; accepted 9 August 1999

Abstract: The purposes of this article are to present an overview of randomized clinical trials (RCTs) and describe some of the methodological problems inherent in using RCTs in nursing research. Many nursing intervention studies are fraught with problems that defy the stringent control criteria required for RCTs, leading to biased estimates of intervention efficacy. Five threats to validity in RCTs are presented, including problems related to (a) differential dropout, (b) random assignment, (c) identifying and maintaining an adequate control condition, (d) nonadherence to research protocols, and (e) assessment of clinically meaningful change. Three strategies are recommended for addressing some of the problems posed by RCTs and improving inference. © 2000 John Wiley & Sons, Inc. 23:79–87, 2000

Keywords: randomized clinical trials; threats to validity of research

When Sir Ronald Fisher (1926) developed the analysis of variance and associated randomized designs, he was conducting agricultural research. Agricultural science as a model for randomized designs is characterized by three features that greatly simplify its implementation and make it a perfect design for nonhuman studies. First, randomizing agricultural samples is easy. Beans identical in strain can be selected and assigned to rows or fields without error. Second, there is no voluntary consent. Short of erosion, beans do not have the option to leave the study. Third, crops are nonreactive and are not prone to placebo effects. For example, a kind farmer will not produce different outcomes than a surly farmer when soil and growing conditions remain constant.

None of these conditions is true in clinical nursing research. When participants are assigned to different psychoeducational conditions, when nurses are asked to implement two different massage protocols with their patients, or when pregnant women are taught a kegel exercise protocol to reduce postpartum urinary incontinence, unexpected situations arise. Participants drop out of the study, staff or participants fail to adhere to well-designed research protocols, participants in the control group react to events outside the study that affect their outcomes, and instruments used to measure change are hampered by ceiling and floor effects. Yet, randomized clinical trials (RCTs) remain the gold standard around which nursing research is taught in our schools, funding agencies allocate resources, and journal reviewers evaluate design integrity. The purposes of this article are to (a) present an overview of RCTs, (b) describe five threats to validity in nursing RCTs that increase the likelihood of making invalid conclusions about intervention efficacy, and (c) offer three strategies that could lead to more valid inferences about outcomes generated by RCTs.

The authors thank Donald W. Fiske for his thoughtful review of this article.
Contract grant sponsor: National Institute for Nursing Research; contract grant number: RO1 NRO4085.
Correspondence to Deborah Gross.
*Assistant Professor.
†Professor.

© 2000 John Wiley & Sons, Inc.
THE RANDOMIZED
CLINICAL TRIAL (RCT)

The RCT is a prospective, clinical experiment designed to test the efficacy of one or more interventions against a controlled condition.Clinical trials have been conducted sporadically in medicine for centuries (Bull, 1959), although its greatest impact has been in the 20th century as it became the model for testing the efficacy of drug therapies, surgeries, and immunizations (Meinert & Tonascia, 1986). Clinical trials were used to test the effectiveness of streptomycin in patients with tuberculosis (Medical Research Council, 1948), methionine for treating infective hepatitis (Wilson, Pollock, & Harris, 1945), and vaccination for the prevention of poliomyelitis (Francis et al., 1955). What is notable about these success stories in the history of clinical trials is that these experiments used (a) chemical treatments, (b) treatments that were administered over a very brief period of time, (c) control groups that could be easily constituted because no cure had been previously offered, and (d) specific clinical events as outcome measures (e.g., infection or death) that required little inference about their significance for health care.

Theoretically, RCTs are true experiments with humans. Thus, the criteria for RCTs are the same as those for any true experiment where the purpose is to establish causal relationships. The first criterion is the inclusion of a control condition that can account for all rival hypotheses that could cause changes in the dependent variable other than those related to the intervention. These rival hypotheses typically include the effects of history, testing, maturation, statistical regression toward the mean, instrumentation error, experimental mortality, and selection biases (Campbell & Stanley, 1963; Cook & Campbell, 1979). The control group is also essential for estimating effect size, which is the difference between the mean effect of the intervention and the mean effect of the control condition. In a control group, participants receive no treatment or a placebo. This is different from a comparison group in which participants receive an alternative treatment, frequently defined by the current standard therapy (e.g., Everett, Patterson, Burns, Montgomery & Heimbach, 1993). The philosophy underlying inference from RCTs is based on the use of control, rather than comparison, groups for estimating effect size (Neyman & Pearson, 1933). This is because control conditions capture a fixed, null condition whereas comparison group effects change over time, depending on the current standard therapy available to participants.

The second criterion for establishing causal relations is the use of an unbiased method for assigning participants to intervention and control groups. Typically, the method employed is random assignment of treatments to individuals. The third criterion is that the intervention and control conditions must be clearly operationalized and designed to maximize experimental variance relative to error variance (Kerlinger, 1986). Although theoretically sound, the feasibility of meeting all three criteria in a nursing RCT is dubious. There are many reasons for this skepticism as described below.

WHAT AILS RANDOMIZED DESIGNS: FIVE THREATS TO VALIDITY

Threats Due to Differential Dropout

As long as voluntary consent is required for participation and participants are given the freedom to withdraw from research studies, RCTs are plagued by the problem of differential dropout. Dropout refers to the decision by some individuals in a RCT to leave the study once they have completed some part of the initial research. If participants dropped out at random, this would be a manageable problem. However, there is strong evidence that participants do not drop out at random (Alterman et al., 1997; Bender, Ikle, DuHamel, & Tinkelman, 1997; McCann et al., 1997; Siddiqui, Flay, & Hu, 1996).

Hollon (personal communication, April 1, 1992) described the phenomenon of nonrandom dropout using an example from research on depression. In this example, the treatment under investigation has both therapeutic benefits as well as annoying side effects. Typically, the most depressed subjects in the intervention group remain because they are willing to put up with the annoying side effects in order to obtain the intervention’s therapeutic effects. Unfortunately, the less depressed subjects drop out of the study because they are not feeling “bad enough” to put up with these side effects. This leads to differential dropout of the healthier subjects in the intervention group. However, the problem is further compounded in the control group. Among controls, the more depressed patients drop out because they are receiving no relief while the less depressed patients remain. This type of differential attrition, which Hollon maintains is quite common in RCTs, leads to nonequivalent groups and inaccurate inferences about the efficacy of clinical interventions. Moreover, differential dropout may leave a study vulnerable to regres-
Threats Due to Problems with Random Assignment

A problem related to dropout is the reluctance of research participants to submit to the essential RCT criterion of random assignment. Random assignment is the equiprobable and independent assignment of treatments to individuals. These treatments usually consist of one or more intervention conditions and a control condition. Because of ethical concerns or the potential for resentful demoralization among participants not receiving the new intervention (Cook & Campbell, 1979), researchers may offer participants an alternate intervention, a placebo, or a delayed intervention. However, so long as there is voluntary consent and ethical codes that require full disclosure in research, participants will make the scientist’s life more difficult by refusing to live by the random numbers table.

An example of the difficulties inherent in random assignment is found in a study by Gross, Fogg, and Tucker (1995). In determining the efficacy of parent training, Gross et al. randomly assigned parents to a 10-week intervention or a waiting-list control group. After the random numbers were selected and applied to the list of participants, each was notified of their assignment. However, one third of the participants who had been assigned to the intervention group requested, instead, to be in the control group. They cited problems such as child care arrangements, busy schedules, and transportation problems that hindered their ability to attend the parent training groups. When the assignments were completed and the groups examined, it was found that parents assigned to the intervention group who remained had the most behaviorally difficult toddlers while parents who dropped out of the intervention group had the least difficult toddlers. In an effort to avoid the dropout problem described earlier, all participants assigned to the intervention group who did not wish to participate in the parent training group were retained in a new group called the “dropout control group.” Nonetheless, earnest efforts to assign participants randomly to treatment conditions had been thwarted.

The difficulties with random assignment in RCTs is not restricted to nursing studies. Participant reluctance to be randomized in cancer chemotherapy and HIV clinical trials has become so problematic that alternative methods for analyzing RCT data are currently being debated in the medical literature. Alternative methods for responding to randomization problems include (a) use of post-hoc severity matching among groups of participants to achieve more meaningful comparisons and (b) redefining unexpected clinical events as pertinent research outcomes to better reflect the clinical course of the disease (Rabeneck, Viscoli, & Horwitz, 1992).

Threats Due to Problems Identifying and Maintaining an Adequate Control Condition

The third source of invalidity in nursing RCTs is associated with the problem of identifying and maintaining an adequate control condition. As indicated earlier, a control group consists of research participants who do not receive the experimental intervention and who constitute the baseline against which the intervention effects are mea-
sured (Polit & Hungler, 1995). The purposes of the control group are to reduce error variance and to prevent rival hypotheses from accounting for the dependent variable (Campbell & Stanley, 1963). Participants in a control condition may be offered nothing or an innocuous intervention that mimics the treatment but excludes its active components (i.e., a placebo).

There are two primary problems with the concept of a control condition in clinical nursing research. The first concerns identifying an adequate control condition. For example, an investigator wants to test the effectiveness of a 30-min educational program on postsurgical recovery taught by the discharge nurse. What type of placebo can encompass the nonintervention characteristics of the educational program? Might a 30-min conversation between the nurse and patient constitute an adequate placebo? Would that placebo be more appropriate than an educational booklet or the same educational program taught by a paraprofessional? That is, is the essential component of this educational program the nurse–patient interaction, the disseminated information, the amount of time the nurse takes to explain the information, the characteristics of the professional nurse, or some combination of these elements? The problem with identifying an adequate control condition is that few theories used to frame nursing interventions are sufficiently well explicated that the inactive features needed to construct a good placebo are made explicit (Gross, Fogg, & Conrad, 1993).

The second problem with identifying appropriate control conditions pertains to the inability to control human beings. By nature, people are reactive. One person may get better because he or she is receiving attention from a health care professional. Another person may improve because a family caregiver moved into his or her home. Still another may get better for unspecified reasons unrelated to the planned intervention. Research has shown that placebo effects can account for improvements in clinical trials (Turner, Deyo, Loeser, Von Korff, & Fodyce, 1994). However, the problem with varying placebo effects in RCTs is that they can mask the beneficial effects of the planned intervention. When the placebo and the planned intervention are both effective, estimated effect sizes shrink. As a result of this variability, interpreting one’s results is difficult and likely to lead to invalid conclusions about intervention effectiveness.

Given the problems involved in designing and maintaining adequate control conditions, many nursing investigators employ one or more comparison conditions. However, as pointed out earlier, inference from RCTs is based on using control groups for effect size. Because comparison conditions change over time, depending on the current standard therapy, effect sizes using comparison groups become difficult to interpret, particularly across studies.

**Threats Due to Nonadherence to Research Protocols**

The fourth threat to validity in nursing RCTs is nonadherence to research protocols. Participants in the intervention condition(s) vary greatly in the degree to which they behave in accordance with the planned protocol. The more complicated and different the protocol is from normal routine, the more likely the individual is to stray. Even when participants are highly motivated to comply, it is unlikely they will adhere precisely to the planned intervention. For example, in a review of federally funded school-based prevention trials, Durlak (1996) reported that most teachers in these studies did not implement the full protocols they had been trained to use. In one study reviewed, only 46% of teachers effectively implemented a school-based “Know-Your-Body” program and in another, 67% of teachers failed to implement a crucial element of a substance abuse program.

There are many possible reasons for nonadherence. Real life may be more hectic and demanding than allowed by the exquisitely detailed but inflexible research protocol. Friends or family may unwittingly conspire to maintain the status quo, even when the status quo is killing the participant. Complex protocols may be too unwieldy to use correctly or too unusual to practice consistently. Regardless of the cause, nonadherence is a pervasive problem for the clinical researcher conducting a RCT, especially when the protocol design requires a change from the study participant’s usual routine. The variability in frequency, intensity, and amount of intervention received when participants do not adhere to research protocols leads to false estimates of intervention efficacy (Yeaton & Sechrest, 1981).

Interestingly, the threat to validity from nonadherence to protocols tends to be built into RCTs since the independent variable is treated as categorical (i.e., the participant is coded as either scheduled to receive or not receive the intervention). The problem of nonadherence as described here suggests that the independent variable should be treated as both categorical and continuous in nature. That is, although the control group ideally receives no intentionally active intervention, the intervention group receives different dosages of
intervention depending on how much each participant adhered to the protocol. For RCTs, the relevant questions should concern both group assignment (categorical data) and dosage received (ordinal data).

Threats to the Assessment of Meaningful Change

The final threat to validity facing the researcher is the assessment of meaningful change. Meaningful change is defined as study differences that are observable and relevant to clinical practice and the intervention participant (Hollon & Flick, 1988). There are four main concerns related to estimating meaningful change in RCTs: (a) ceiling and floor effects often associated with continuous health outcome measures, (b) differential reactivity to measures, (c) determining the appropriate time intervals for measuring health change, and (d) identifying the amount of health change that is clinically significant. Clinical significance, in this context, refers to the amount of health change that has a significant effect on the participant’s health.

Selecting outcome measures that are sensitive to differences between the experimental and control groups is essential for estimating intervention efficacy (Stewart & Archbold, 1993). Ideally, researchers should use measures with strong reliability and validity and have evidence from pilot work that the measures are sensitive to change. However, some measures that are relevant to the intervention and have good construct validity represent characteristics that are not well distributed within the population of interest. If participant baseline scores are clumped at one of the extreme ends of the scale, there may be no room for improvement, resulting in attenuation of intervention effects. This phenomenon, referred to as either ceiling (clumping at the high end of the scale) or floor (clumping at the low end of the scale) effects, is common in studies on prevention, health promotion, and chronic illness (Eastman, Young, Fogg, Liu, & Meaden, 1998; Shaffer, Philips, Garland, & Bacon, 1992).

An example of floor effects is found in a study by Johnson (1996), who examined the effect of an audiotaped patient education intervention on coping and disruption of usual activities among patients receiving radiation therapy for prostate cancer. Johnson found that baseline scores for disruption of activities were skewed toward zero because most patients reported no preintervention disruption in their activities. Moreover, patients continued to report few disruptions in their usual activities during radiation therapy. This floor effect on disruption of usual activity reduced the intervention’s apparent effect and may have resulted in an underestimation of the intervention’s clinical effectiveness.

Another problem associated with the assessment of meaningful change is differential reactivity to measures. This occurs when participants in the intervention and control conditions react differently to the research measures because they have had different experiences with one or more aspects of the research. This bias is different from “experimenter effects,” when researchers communicate expectations to participants about how the participants should respond to the experiment. Differential reactivity to measures results from systematic changes from pre- to postintervention in how willing participants are to respond honestly or naturally during assessments.

An example of differential reactivity to measures is from a study on the effectiveness of parent training for reducing child behavior problems among African American and Latino day care children in low-income communities in Chicago (Gross, Fogg, Webster-Stratton, & Grady, 1999). Typical of low-income urban families of color (Anderson, 1993; Gamble, 1993), the study parents were wary of researchers and particularly suspicious of whether the data would be shared with government agencies associated with child protective services, public aid, or immigration. However, suspiciousness among the intervention parents appeared to abate by the end of the intervention as a result of their positive experiences with the parent training program and as evidenced by their high consumer satisfaction scores.

Throughout the program, intervention parents came to trust the members of the research team and began to disclose more information about themselves and their children. According to the data collectors, intervention parents appeared more open and honest at the postintervention assessments than they had been previously (Julion, Gross, & McLaughlin, in press), disclosing more problems than they had reported at preintervention. Thus, intervention parents may have underestimated their problems at preintervention and more accurately reported their problems postintervention, yielding little or no change in parent-reported child behavior problems. Nonsignificant effects would suggest that the parent training program had been ineffective in reducing child behavior problems. However, parent satisfaction scores, observations of parent–child interactions by raters blind to conditions, and behavioral checklists completed by day care teachers revealed that significant improvements in child behavior had
ocurred. This pattern in results indicates that the parent-report measures may have been affected by systematic differences in participants’ experiences of the research by condition. It also highlights the importance of multiple measures and informants when assessing intervention efficacy.

Assessing meaningful change also is hampered by the problem of when to assess change. The most basic experimental design, the parallel RCT, employs pretest and posttest measures that are examined using an analysis of variance model (Moody, 1990). Timing of the outcome measurement is based on predetermined endpoints tied to the initiation and conclusion of the intervention. However, it is possible that the effects of the intervention are not fully actualized until sometime after the posttest is over, when the individual has either had the opportunity to incorporate the intervention into their daily life or the body’s physiology has fully responded. In effect, there are three aspects to treatment response: (a) the length of time until the intervention produces clinical improvement, (b) the rate of improvement, and (c) the maximum level of functioning the participant can attain using this particular intervention. If the researcher assesses outcomes too early, the full effectiveness of the intervention will not be detected and the maximum level of functioning that can be attained from the intervention will never be known. If the researcher assesses outcomes too late, it will appear that the change was less rapid and, hence, the intervention less effective than is the actual case. In either case, intervention effectiveness is underestimated because inappropriate intervals were selected for detecting change.

Finally, as LeFort (1993) has pointed out, there is little consensus regarding what constitutes clinically significant change. Although many approaches have been applied, it remains unclear just how much change or what kind of change is clinically important in nursing intervention studies. Oftentimes, statistically significant change obtained from a RCT is accepted as evidence that the intervention “worked.” However, there has been little research on what constitutes significant clinical change in nursing research.

The need to identify useful markers of clinically significant change is also apparent in medical research. As one physician noted (Feinstein, 1983), cancer treatment outcomes defined by tumor size, white blood count, and survival time does not indicate whether a patient is functional and vibrant or miserable and vegetating. Many important intervention benefits are difficult to capture in numeric values, pointing to the potential of using qualitative outcomes in RCTs for understanding meaningful change (Sandelowski, 1996). The problems of differential dropout, inability to randomly assign participants to groups, inability to identify and maintain adequate control conditions, nonadherence to research protocols, and difficulties assessing clinically significant change suggest that the results of RCTs should not be accepted without careful assessment. Each of these problems introduces systematic bias that makes it difficult to validly interpret the results of a RCT. Given the complexities of nursing interventions, clinical scientists need to explore alternative strategies that are sensitive to the human variations that threaten RCT validity. We offer several ideas for future research.

**STRATEGIES FOR IMPROVING INFERENCE**

**Adding a Dose-Response Model to RCTs**

The purpose of the dose-response strategy is to examine the impact of increasing intervention dosage on the dependent variable. The hypothesis underlying this strategy is that if the intervention is effective, dosage should be systematically related to improvements in outcome(s). The greatest advantage to adding a dose-response model to RCTs is that nonadherence to the research intervention becomes part of the study design. A second advantage is that the problem of measuring change secondary to floor and ceiling effects are potentially reduced with a dose-response model because the variability in dosage adds fidelity to the measurement of change.

Dose-response models are not new. They have been used extensively in the past 50 years to study the efficacy of varying drug dosages to treat illness and to determine the optimal dosage for achieving clinically meaningful change (e.g., Rosenberg & Arling, 1944). Dose-response models were recently used in nursing studies examining the effectiveness of a self-help program for breast cancer survivors (Sidani, 1998) and of a parent training program (Tucker, Gross, Fogg, Delaney, & Laporte, 1998), both compromised by variable participant attendance.

The degree to which dose-response models enhance the researcher’s ability to evaluate intervention efficacy is directly related to how dosage is conceptualized and operationalized in a nursing study. For example, a study on the effectiveness of a cancer support group on survivor immunity may require several types of dosage measures such as number of sessions attended, frequency of verbal
contributions made in the support group, and frequency of contact with other group members outside the support group sessions. As with any intervention study, a good theory is essential. The theory that explains why the intervention should work is also the theory necessary for explaining which components should be addressed in the operationalization of dosage.

Including Qualitative Outcome Measures for Assessing Meaningful Change

In RCTs, intervention success largely rests on the achievement of statistical significance between intervention and control groups using quantitative measures. As described above, there are numerous problems in using quantitative measures for assessing clinically meaningful change in RCTs, leading to faulty conclusions about intervention efficacy.

Qualitative data can contribute substantively to understanding the effectiveness of clinical interventions by providing information about intervention utility and significance as seen through the eyes of the participants (Sandelowski, 1996). For example, a RCT targeting stress reduction in caregivers might yield no statistically significant change in total stress scores even though participants experienced meaningful changes in their lives following the intervention. Without methods that address the within-group differences that attenuate effect sizes, researchers may inaccurately conclude that their interventions are ineffective. The addition of qualitative assessments for evaluating the outcomes of RCTs is likely to improve our ability to more fully assess intervention efficacy based on change that is meaningful to the participants.

Developing Participant-Centered Methodologies

Estimation of RCT effectiveness could be enhanced by developing more “participant-centered” methodologies. Participant-centered methodologies are defined as research strategies that are consistent with participants’ values and the contexts in which they live. The advantages of such methods are that they increase the likelihood that participants will remain in RCTs and adhere to protocols because the research will be consonant with their daily experiences. In addition, participant-centered research will increase the social validity of the research because improvements that result from the intervention are more likely to be meaningful to the target population.

For example, focus or advisory groups comprised of members of the target population might be used to develop or modify research protocols so that they are more consistent with participants’ daily activities (Hughes & DuMont, 1993). Similar groups might be used to identify outcomes that are particularly valued that would represent meaningful improvements. Hatch, Moss, Saran, Presley-Cantrell, and Mallory (1993) recommend that researchers and participants identify “dual agendas” for determining research outcomes. In their study on aerobic exercise on cardiovascular fitness, researchers funded by the American Heart Association were interested in the psychosocial dynamics of getting African American women involved and committed to a cardiovascular fitness program. The women participating in the study were mainly interested in feeling better and looking better. As a result, additional outcomes were added to the research protocol, including the women’s abilities to reduce their waist size. Participants eventually organized a fashion show at local churches to show off their “outcomes.”

CONCLUSIONS

RCTs require a great many resources to implement. They are complex, expensive, and highly demanding of participants, researchers, and supporting agencies. Whereas crops can be sorted and treated in a controlled fashion, people are not so tractable. Participants drop out of RCTs, fail to adhere to their assigned regimens, and react to influences that are beyond the researcher’s control. What emerges, according to Salsburg (1992) is “a complex collection of data filled with random noise, loosely organized around a fiction called the protocol” (p.4). While the RCT may still be the best available method for determining intervention efficacy, it needs work. RCTs need to be supplemented with methods that address the human elements of clinical research.

REFERENCES


