Pitfalls of Multisite Randomized Clinical Trials of Efficacy and Effectiveness

by Helena Chmura Kraemer

Abstract

This article discusses common pitfalls of multisite randomized clinical trials of efficacy and effectiveness. Issues considered include (1) premature initiation, (2) ineffective study structure, (3) too much or too little communication among researchers, (4) neglect of site differences, (5) use of “naked” p values, (6) premature closure, and (7) overreliance on the interpretation and memory of individual researchers. If future researchers are aware of these common pitfalls, they may be able to avoid them.

Keywords: Multisite studies, randomized clinical trials, research planning, research organization.


Progress, far from consisting in change, depends on retentiveness. Those who cannot remember the past are condemned to repeat it.
—George Santayana (1905/1988)

In recent years there has been much greater emphasis, particularly in mental health research, on multisite randomized clinical trials (RCTs), particularly in areas concerned with the effectiveness rather than the efficacy of treatments. Much has been written about these two important and controversial related trends: (1) the focus on efficacy or effectiveness and (2) use of the multisite RCT to address such issues.

Multisite RCTs

Meinert (1986) defines a multisite RCT as one having three characteristics: (1) the study must involve two or more clinical sites and their separate staffs; (2) all sites must follow a common treatment and data collection protocol; and (3) one center is charged with accruing, processing, and analyzing the data from all the sites. Under this definition, one site with various subject recruitment sources does not qualify as a multisite RCT but continues to be a single-site RCT, stratified perhaps by recruitment source. A study involving multiple sites with different recruitment, treatment, or measurement protocols does not qualify as a multisite RCT but as multiple collaborating single-site RCTs.

There are good reasons why researchers and funders stage multisite RCTs. First, multisite RCTs can generate larger sample sizes, with greater power to test hypotheses and more precise estimation of population parameters. This advantage is particularly important in the study of low-prevalence disorders, where it may be impossible to generate a large enough sample size at any one site.

Second, the findings of multisite RCTs are more generalizable than the findings of single-site RCTs. Sites in multisite RCTs often have subjects with varied sociodemographic and clinical characteristics and unique treatment and assessment staffs (even under protocols uniformly defined and implemented across the sites). The effect of treatment may differ from one site to another (site by treatment interactions) even when the treatments are uniformly delivered and evaluated (Infant Health and Development Program 1990). Understanding how circumstances affect treatment efficacy—vital to preventing overgeneralization of conclusions—is impossible in any one single-site study and difficult even in multiple single-site studies.

Third, and most important, multisite RCTs can help resolve the most contentious conflicts in a field. Decisions
governing research protocol formulation and delivery are often consciously or unconsciously influenced by the beliefs of the investigators. That pharmacotherapy fares better than psychotherapy, for example, when evaluated at a site more committed to pharmacotherapy (while the reverse is true at a site more committed to psychotherapy) should not be a surprise. When the two sites are brought together to design and implement one protocol to compare the two treatments, the treatment protocols for pharmacotherapy and psychotherapy are likely to reflect the best available protocol for each, and to be better and more uniformly implemented at every site. When such sites must agree on the appropriate outcome measures, the selected outcomes are more likely to reflect the best each treatment is able to offer. As a result, contentious questions concerning the effects of treatment are more likely to be resolved in a multisite study than in even high-quality multiple single-site studies. In any case, with many experienced researchers involved in the design and execution, multisite studies are likely to be better designed and implemented and the results more carefully and expertly reported than single-site studies.

On the other hand, errors occurring in a multisite RCT are likely to be more serious and lasting. Given the salience of multisite RCTs and their great cost, an error may not be detected for some time. It is difficult for a single-site study to challenge the conclusions of a multisite RCT. In fact, if a great deal of money is spent on a multisite RCT addressing a particular issue, fewer funds may be available for subsequent single-site RCTs addressing that same issue. These factors delay, and perhaps even prevent, the detection and correction of errors in multisite RCTs.

Efficacy Versus Effectiveness

Another topic currently debated is that of efficacy versus effectiveness. Although frequently so used, these terms are not antonyms (Hoagwood et al. 1995). Instead, efficacy and effectiveness are opposite extremes on a complex multidimensional continuum of decision making in research design. This section discusses a few examples to illustrate how focusing on efficacy or effectiveness affects study design.

The Population Sampled. On the efficacy side, very stringent inclusion and exclusion criteria are applied. For example, in studying subjects with schizophrenia, one might exclude all those with any serious comorbidity, thus excluding most schizophrenia patients. Because the Food and Drug Administration (FDA) does not view its mandate for approval of drugs as requiring demonstration that they work for a typical or representative subject having the disorder, FDA phase III studies most frequently lie at the efficacy end of the spectrum. On the effectiveness side, people are excluded only when there are empirical or ethical reasons to do so (e.g., refusal to give informed consent, possibility of injury from one of the treatments in the study). This effectiveness approach creates a very heterogeneous sample requiring larger sample sizes, but the results are more likely generalizable.

The Delivery of Treatment. On the efficacy side, treatment may be delivered only by those expert in treatment of that particular disorder at tertiary care centers that specialize in the treatment of that disorder, with resources and equipment unavailable to many outside such a setting. On the effectiveness side, treatment is delivered by physicians who ordinarily would treat that disorder and are well-motivated and well-trained but not necessarily the world’s experts. They use resources and equipment available to the typical physician’s office, clinic, or hospital.

Outcome Measures. On the efficacy side, high-technology, costly, and esoteric outcomes are often preferred. The outcomes give researchers greater insight into the etiology and process of the disease, often involving costly medical tests such as genotyping, nuclear magnetic resonances, or interviews lasting many hours per patient. Given a choice between a “good” measure (i.e., a sensitive and highly reliable one) and a “right” measure (i.e., a valid one, meaningful to patients and their physicians), the preference is often given to “good” measures (Tukey 1979). On the effectiveness side, the reverse is true. There one sees measures of patient satisfaction, cost, quality of life, patient perception, direct clinical observation, and so forth—measures meaningful to patients and their physicians but often hard to measure objectively and reliably.

Choice of Control Group. On the efficacy side, the control group is likely to involve an inert placebo. On the effectiveness side, the control group is more likely to involve usual care or the current standard of treatment for the disorder.

Points on the Continuum. Historically, studies on the efficacy side have been expected to be conducted under the full rigor of RCT methodology, but studies on the effectiveness side have often been observational, with little or no control of selection or evaluation biases. Consequently, it has sometimes been claimed that a further distinction between efficacy and effectiveness trials is that on the efficacy side, one sees “clean,” rigorous studies, while on the effectiveness side, one sees “dirty,” nonrigorous ones. If that were true, there would be a distressing situation in which only the right answers to the wrong questions
(extreme efficacy studies) and the wrong answers to the right questions (extreme effectiveness studies) were found. Instead, the principles of RCT methodology should apply wherever a study is positioned on the efficacy/effectiveness continuum. Most studies are neither totally on the efficacy end nor totally on the effectiveness end; rather, they are at some midpoint that reflects the current state of knowledge about the treatments involved. In the following section, no distinction will be made between efficacy RCTs and effectiveness RCTs. However, since multisite RCTs are more likely to be required on the effectiveness side, these pitfalls may apply more strongly to effectiveness RCTs.

Pitfalls

Any pitfall that pertains to a single-site RCT also pertains to a multisite RCT. Researchers and biostatisticians who have worked in multisite RCTs are always happy (at least privately) to describe the various traps that they have fallen into or just missed or the disasters that have occurred or were narrowly averted. Mosteller and colleagues (1980) suggested that there should be an award for the most bizarre disaster in an RCT, in hopes that widely disseminating information about such experiences would help prevent future research studies from falling into exactly the same trap. Consistent with that suggestion, the pitfalls here discussed reflect some of my own experiences and regrets.

Premature Initiation. Multisite RCTs are sometimes proposed when single-site RCTs and other studies have not generated much good information about an issue—when questions such as what outcome measures are suitable, what designs are likely to succeed, and what effect sizes are necessary for power calculations must be based on guesses rather than data (Meinert 1980). This lack of information creates a serious problem.

Interventions of the type evaluated in RCTs often originate either in clinical observation, perhaps in case histories, or as a byproduct of basic science research (FDA stage 1). Early in the development of a treatment, there are uncontrolled or nonrandomized single-site studies with small samples, even single-subject studies, none of which are likely to establish the efficacy or effectiveness of a treatment definitively (stage 2). However, such studies are both necessary and valuable for refining the treatment protocol, identifying the appropriate inclusion and exclusion criteria, and helping in the selection of outcome domains and their measurement. The studies may also be used as a basis of discussion of what a clinical significant effect size might be in each domain and to yield preliminary estimates (typically optimistic at this stage) of the effect size for the treatment under consideration, to be used in power calculations. All these insights and decisions are necessary to the design of any RCT (stage 3).

In stage 3, ideally many single-site RCTs are done, many with research protocols (sample selection, choice of control group, protocol of treatment, choice of outcome domains and measurements) differing from each other. As progress is made through this stage, ideally the results of the RCTs could be tracked sequentially via meta-analysis (Lau et al. 1992). One of four decisions would follow each sequential meta-analysis:

1. The RCTs done to date are inconclusive, and more single-site RCTs are needed to better define the research question or research strategies.
2. The RCTs done to date are conclusive. The effect of treatment is not only statistically significant but of a magnitude to be of clinical or policy significance as well. No further studies are needed.
3. The RCTs done to date are conclusive. The effect of treatment, statistically significant or not, is of a magnitude unlikely to be of clinical or policy significance. No further studies are needed.
4. The RCTs done to date are inconclusive, and there is sufficient heterogeneity among their results, incongruity in the procedures used, or inadequate sample sizes that further single-site RCTs are unlikely to answer the question. A multisite RCT is warranted and justified.

By the time it is clear that a multisite RCT is needed, a great deal of background should have been established as a basis for designing and implementing the multisite RCT. That carries the implication that the multisite RCT should be more efficiently designed, offering greater power at less cost, than any single-site RCT. To propose a multisite study when the question could be resolved in single-site studies is wasteful. However, to propose a multisite RCT before the research question and the approaches to take are clear from an accumulation of single-site RCTs may result in a poorly designed and inadequate multisite RCT that not only would waste a great deal of time, effort, and funding that could have been better invested elsewhere but also might mislead the entire field for some time.

Ineffective Study Structure. It is uncommon to spend much time and effort considering the organizational structure of a single-site RCT. That is certainly not what biostatisticians and researchers are trained to do. But multisite studies are different. Without careful attention to structure, a multisite study’s scientific quality may be undermined, and the study may be more expensive, difficult, and frustrating. Much of what is said here follows Meinert’s (1986) suggestions quite closely.
First of all, the buck must stop somewhere. There must be one person, the study's principal investigator (PI), who coordinates the study as a whole, adjudicates controversies, makes and enforces the tough decisions on issues related to science as well as to personnel and the budget, and maintains the documentation of the conduct of the trial. The study PI may be self-appointed (typically the person who organized the multisite RCT), appointed by the funders (by the National Institute of Mental Health, perhaps), elected by the site PIs of the study as a permanent leader, or rotated among the site PIs either in turn or by election over the course of the study. The worst possible situation in a multisite RCT is one where discussions never end, where disputes are never resolved, and where decisions are unclear, unenforced, or inordinately delayed, all because of the lack of a responsible and effective leader.

The executive committee is a group of people responsible not to any one site but to the study as a whole. Its members compile, update, and distribute documentation (e.g., protocols from treatment and measurement); implement study decisions; arrange meetings; monitor and administer the overall study budget; and coordinate communication among sites.

The site directors or site PIs are responsible for overseeing the participation of each site in the multisite RCT. Ideally, the site PIs represent diverse scientific viewpoints. Thus, in a multisite RCT comparing pharmacotherapy and psychotherapy, some PIs should be advocates of pharmacotherapy and some of psychotherapy, and both groups should have participated in the multiple single-site RCTs that produced the conflicting results. This balance may produce good science but be difficult to achieve. Some investigators who are superbly qualified to conduct a single-site RCT are not adept at the interactions and negotiation necessary for a successful multisite RCT. To include such investigators as site PIs in a multisite RCT may compromise both the scientific quality of the study (e.g., because of lack of compliance with the common protocols at these investigators' sites) and the efficient and timely conduct of the RCT (e.g., because necessary decision making may be unduly impeded). In structuring a multisite RCT, it might be advisable to have a funded trial period during which many more potential site PIs are involved in the early planning of the study than can be involved in the study itself. The sites would be selected at the end of this time period rather than before it. Site selection would then (as now) be based on consideration of factors such as the availability of suitable subjects at the site, the ability of the site to implement the study protocol, and the skills of the site PI. Selection would also (as does not now seem true), however, be based on consideration of the ability of each potential site PI to interact productively with the other site PIs to resolve the conflicting views bound to arise. Each site PI must be committed to finding the correct answer, even if that answer goes against his or her own previous research findings.

One of the requirements of a multisite RCT is that there be central data processing, which is headed by a data base manager (DBM). The DBM, who may or may not also be the study biostatistician, must have expertise in data management, data entry, data checking, data retrieval, implementation of data analysis, and so forth. The study biostatistician addresses conceptualization, design, measurement, analysis, and interpretation but not necessarily implementation. Clearly, if the DBM and the study biostatistician are different people, the interactions between them must be excellent, for one can easily subvert the efforts of the other. Moreover, it is not a good idea to have only a DBM who is not a well-trained and experienced biostatistician, or only a biostatistician who is not a well-trained and experienced DBM.

There must also be a research steering committee (RSC). This group makes all the decisions concerning the scientific conduct of the study. This committee should include at least some or all of the site PIs, the study PI, the study biostatistician, and the DBM. The RSC may include members not otherwise connected with the study, experts in relevant areas not well-represented in the multisite group. Members of the RSC should not regard themselves as advisory but as actual participants in the multisite study, with responsibilities to the study beyond attending meetings. This group would be expected not only to do the library or laboratory work necessary for sound decision making but to provide the written documentation for such decisions. Finally, outside experts may be invited to sit on an ad hoc basis with the RSC to deal with particular research questions.

Often there is a data safety and monitoring board (DSMB). Ideally, the DSMB is an external group that reports not to the sites or to the study PI but to the sponsor of the multisite RCT. Unlike the RSC, this is an advisory group whose members do not actively participate in the multisite RCT. The DSMB's members try to take an objective, professional view of the study. They are invaluable in that they often anticipate the objections, questions, or reservations of future reviewers and readers while the study can still make modifications. As the board's title implies, the DSMB is often charged with taking the perspective of the study subjects and considering any ethical issues that might arise. This would be the group likely to recommend to funders that a study be ended prematurely for safety reasons.

Except for the DSMB members, people involved in the study may play several roles. There should be one study PI, but the size and composition of the other groups
Too Much or Too Little Communication Among Researchers. Even in the relatively recent past, it would have been unthinkable to worry about too much communication. However, with ready access to the telephone, conference calls, teleconferencing, e-mail, fax, voice-mail, express mail, U.S. mail, and so forth, too much communication can be a serious problem. A researcher may spend many more hours sorting through unnecessary and misleading communications related to a study than actually working on the study.

Here is an extreme example: Suppose one participant in a multisite RCT is unsure of the exact protocol for entering a particular subject, so she sends an e-mail to 20 other participants asking for clarification. Each of those 20 contacted may respond. Some will respond with the right answer, others will offer the wrong answer, and others will respond that they too are unsure. Each of these 20 messages may be copied to all 20 others. At this point, there may be some 400 messages, some of which carry misinformation. Then there must be messages identifying and correcting the misinformation as well as providing the correct answer. Some participants may later remember the wrong answers more clearly than the right answers.

There should be a "chain of command" for study communications. Any question concerning the protocol for the study should be referred to the study PI or to the head of the executive committee. If the issue is already covered in the existing protocol, the answer (and a reference to the protocol documentation) should be provided directly to the inquirer, perhaps with copies sent as a reminder to others. If the issue is not already well-covered, the query should be referred to the head of whichever committee would be the responsible one: to the executive committee for issues concerning organization personnel and budget, to the RSC for questions about the scientific conduct of the study, and to the DBM for questions about data accrual and management. It is then the responsibility of the head of that committee to get the issue settled expeditiously by the committee and to provide an update to the protocol for distribution to all sites.

The more familiar and still valid complaint is that of too little communication. In a single-site study, decisions can often be made "on the fly" without seriously impairing the scientific quality of the study. In a multisite RCT, because the quality of the study depends on the same decision being implemented at all sites, this on-the-fly approach is simply not possible. Each decision made that affects the conduct of the study, and every amplification or modification to such a decision, must be documented and distributed to all sites (Kupfer et al. 1994). When this approach is not used, there is often "slippage" among the sites, growing greater as the study progresses, until one ends, not with the planned multisite RCT but with multiple inadequate single-site RCTs. Since the prime reason for proposing a multisite RCT is that it is unlikely that multiple single-site RCTs can answer the question, this is a disastrous result. Communication channels must be well-defined, open, and effective.

Moreover, decisions must be made and communicated early about access to the common data, about authorship, about credit, and so forth. Misunderstandings about such important issues may well endanger the study later, for example, if one site's results are prematurely published or if resentments arise that affect compliance with the common protocol.

Whether or not ancillary ("piggyback") studies are allowed must also be settled and communicated early. Such studies address issues different from those for which the multisite RCT is designed but use either some or all of the same subjects or data as the multisite RCT. Such ancillary studies may be conducted at only one site, several sites, or all sites. The studies often reflect the individual interests of some site PIs, and their results appear under the authorship of the personnel actually and directly involved in the ancillary study, not those of the whole multisite RCT.

Clearly, any ancillary studies that threaten the success of the study as a whole must be discouraged. For example, the burden of data collection can result in patient dropout and often is a source of increased unreliability of measurement. Dropout compromises both sample representativeness and comparability of randomized groups as well as power. Unreliability also attenuates power (Kraemer 1991). The RSC would be ill-advised to approve an ancillary study that requires collection of much data beyond that required in the multisite RCT, for that may substantially increase the burden and compromise the validity and power of the multisite RCT.

On the other hand, a comprehensive baseline (prerandomization) data set is necessary to the multisite RCT and can often simultaneously serve as a cross-sectional epidemiologic data base. Complete characterization of the population from an ancillary study done with baseline data is often an advantage later to presenting the conclusions of the multisite RCT. Sometimes the control group
of a multisite RCT can be used for longitudinal epidemiologic or methodologic studies, again with possible advantages to the multisite RCT. The RSC should not only approve but encourage such studies.

The credit each investigator in a multisite RCT receives is diluted by the number of investigators involved, even for the most important and well-done RCT (Meinert 1980). Junior faculty members in particular may not be given credit commensurate with their effort, time, and excellent contribution, and this may be deleterious to career advancement. Ancillary studies on which the individual is the senior author can ameliorate this unhappy situation in facilitating studies with a scope much wider than a single-site study of any kind. The decisions the RSC must make with regard to ancillary studies are difficult and complex, meriting early and complete consideration and communication to all study participants.

Finally, the issue of whether and when the study data will be archived for public access should be decided early. To do so benefits the study, since the documentation must be clear, not only for the participants in the study but for all future users of the data set. Also, when study participants are conscious that every decision, analysis, and conclusion may be double-checked by others, greater care is taken in the decision making and its documentation. This directly improves study quality. Moreover, because multisite RCTs have greater salience than single-site RCTs and because the full value of the multisite RCT data set may be much greater when eventually opened for the use of nonparticipants, there is typically good reason to decide on such public archiving. In doing so, of course, suitable attention must be paid to confidentiality issues.

Neglect of Site Differences. All the pitfalls discussed so far relate to planning, organization, or implementation of the research enterprise. This section discusses an analytic issue: neglect of site differences.

Table 1 shows a hypothetical case with two sites and two treatments. At each site, the population effect size (the standardized mean difference, \(d\)) is exactly the same: 4.0. There is here no site \(\times\) treatment interaction, but there are clear site differences. According to the table, 50 percent of subjects are assigned to treatment and 50 percent to control, but the number per site per group differs.

A frequent justification for ignoring site differences is that no "statistically significant" site \(\times\) treatment interaction was found. There are two errors in such a justification. First, finding no "statistically significant" interaction does not prove the null hypothesis that there is no interaction. The sample size necessary to have adequate power to detect an interaction is typically many times larger than that to detect a main effect of treatment (Cohen 1988). A nonsignificant interaction may still, if ignored, bias treatment effects. Second, a fact to be shown in the current example, a nonexistent interaction does not protect against bias when site differences are ignored.

If one ignores site differences completely in the analysis for the population in table 1, the mean response for the treated subjects is somewhere between 50 and 90, and the mean response for control subjects somewhere between 10 and 50, depending on what proportion of the subjects in each group are at each site. Moreover, the within-group standard deviation lies somewhere between 10 (when all of one group is at one site) and 22 (when half of the group is at each site). The sample effect size, therefore, varies between 0.0 and 8.0. It is very unlikely that one would hit the right value of 4.0 by chance. Figure 1 shows sets of points that yield varying effect sizes, all from the data of table 1.

The simple answer, of course, is that site effects and site \(\times\) treatment interactions should always be considered in the analysis of the results of a multisite RCT. Ignoring site differences can produce a statistical artifact called Simpson’s Paradox (Blyth 1972; Bickel and Hammel

<table>
<thead>
<tr>
<th>Site</th>
<th>Treatment</th>
<th>Control</th>
<th>Effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Site 1</td>
<td>90 ± 10</td>
<td>50 ± 10</td>
<td>4.0</td>
</tr>
<tr>
<td>Site 2</td>
<td>50 ± 10</td>
<td>10 ± 10</td>
<td>4.0</td>
</tr>
</tbody>
</table>

Figure 1. Effect sizes comparing treatment (T) and control (C) groups when the true effect size at each site is 4.0 and site differences are ignored, for differing percentages of each group at each site.
relations), but the factor may be reported as a strong risk factor, sometimes even proposed as causal, for the disorder. The true correlation between factor and disorder over all the subjects sampled (muddled control population, and a putative risk factor is correlated with a disorder from a low-risk or test population, another sample is drawn from a high-risk or test population, in which one draws samples from a number of different populations (in RCTs, sites) and evaluates the association or correlation (in RCTs, effect sizes) between two variables (in RCTs, treatment choice versus response). Two facts are known about this situation:

1. The correlation between subjects within one population may be very different from that in another. Each such correlation is called an “individual correlation.”

2. If the means of the two variables in each population (in RCTs, the percentage of subjects in the treatment group, T, at each site versus the average response at that site) are correlated across the populations, this correlation bears no necessary relationship to any of the individual correlations. It may be highly positive when every individual correlation is zero or vice versa. This correlation is called the “ecological correlation.” To assume that the ecological correlation somehow estimates the individual correlations is what is referred to as the “ecological fallacy.”

In this situation, if one simply threw all the paired observations together, ignoring which population each observation was drawn from, the resulting correlation (the “muddled correlation”) would be a weighted mixture of the individual correlations and the ecological correlations (Kraemer 1978). How much each correlation would influence the muddled correlation depends on what proportion of the total variance of the observations results from within-population and between-population effects. As a result, it is quite possible to have a muddled correlation equal to zero when each individual correlation is large and positive (as in the above illustration), and it is quite possible to have a muddled correlation that is large and positive when every individual correlation is zero.

The problem of muddled results is best known, not in the context of RCTs, but in that of epidemiologic studies, where it is known as Berkson’s Fallacy. This is a situation in which one sample is drawn from a high-risk or test population, another sample is drawn from a low-risk or control population, and a putative risk factor is correlated with a disorder over all the subjects sampled (muddled correlation). The true correlation between factor and disorder may be zero in every subpopulation (individual correlations), but the factor may be reported as a strong risk factor, sometimes even proposed as causal, for the disorder because there is a high ecological correlation that determines the muddled correlation. Clearly, such reports can seriously mislead the search for causes of disorders. However, the problem is no less important in multisite RCTs, where reporting muddled results (i.e., ignoring site differences that may be important) may mislead clinical decision making and future research.

Use of “Naked” p Values. If a multisite RCT is not initiated prematurely, there should be enough background studies to suggest strongly that the null hypothesis of random differences between treatment and control is unlikely. To spend time and effort merely to reject such a null hypothesis seems wasteful (Meehl 1967). Instead, the study should attempt to accurately estimate the effect size of the treatment, to establish that this effect is large enough to support claims of clinical or policy significance. Statistical significance is necessary, but not sufficient, to such a claim. For this reason, reporting the results of a multisite RCT only in terms of test statistics or p values should be discouraged. Many a statistically significant result is of little clinical or practical importance. Effect sizes that are informative to the clinical and policy audience, along with their confidence intervals, should be the primary focus of an outcome paper of a multisite RCT (Borenstein 1994).

Premature Closure. Just as multisite RCTs are sometimes initiated prematurely, they are sometimes closed prematurely, before sufficient analysis is done to fully understand the results and their repercussions.

Each person in the population sampled in any RCT has an individual “causal effect of treatment,” defined as the difference between that person’s response if given the treatment and that same person’s response if given the control condition (Rubin 1974; Holland 1986). Clearly this individual effect size cannot be estimated, for no individual can be given both the treatment and control conditions simultaneously. (And giving them in sequence mixes in time effects, carryover effects, and so forth.) The effect size in an RCT estimates the average of these individual effect sizes over the individuals in that population.

There is nothing in the averaging process that precludes variability among the individuals in that population. With the current overemphasis on null hypothesis statistical testing (Dar et al. 1994; Cohen 1995; Harris 1997; Hunter 1997; Shrout 1997), the exploration designed to detect and understand such variability has been discouraged. However, in absence of such exploration, a treatment is liable to be recommended and used for many subjects in the population for whom it does no good and perhaps even does harm (moderating variables) (Baron and Kenny 1986). In absence of such exploration, how and why a treatment works or does not (mediating...
variables; Baron and Kenny 1986) remains unknown. As a result, information useful to improving treatment effects is simply ignored (Kazdin and Weisz 1998).

**Overreliance on the Interpretation and Memory of Individual Researchers.** A recurrent theme here is the importance of a uniform protocol that is fully documented. In a multisite RCT, the same protocol must be followed at all sites. The formulation, amendment, and circulation of the research protocol are fundamental duties of the executive committee. The documentation to be included in the research protocol is a fundamental duty of the RSC. The implementation of the same protocol at all sites is a fundamental duty of the site PIs.

In a multisite RCT, one would expect very strong, and often competing, views on a variety of issues. Such issues must be discussed and decisions must be made, documented, and adhered to at every site. Problems arise when a crucial decision is made based on personal opinions of one or a few people and there is no documentation supporting that decision. The following two examples illustrate this problem.

In one multisite RCT, an RSC member strongly argued for the use of a particular short, noninvasive test claimed to be a reliable and valid literacy test. The executive committee located the documentation for the test, which showed that in fact the proposed test was not a valid test of literacy. In another case, one researcher in a multisite study insisted on the use of a Latin Square design to deal with the carryover and interactive effects of the repeated tests per subject. An outside statistical consultant provided documentation that indicated that the Latin Square design was contraindicated whenever such carryover and interactive effects were expected. Both researchers making these proposals and arguing strongly for them were experts in their fields, but they simply misremembered or misinterpreted what they had read.

For that reason, when an important decision is made in the design and conduct of a multisite RCT, it is important that the decision be based not on personal opinion, memory, or interpretation of even the most expert and experienced researcher or on the most persuasive argument, but on documentation. Interacting well and productively within the multisite RCT setting requires that a participant be willing to provide documentation supporting his or her own views, to thoughtfully review and process documentation supporting contrary views, and then to accept the final decision of the group.

However, in many cases, there are several “right” answers on an issue, each of which can be documented. If there are strong competing views, a good strategy is to ask the strongest advocates of each view to write a position paper supporting the view, taking into consideration the documentation on all sides. The decision is then based on a review of the position papers. Such position papers become part of the documentation, showing the care with which the decision was made and explaining the scientific basis of the decision. Frequently, the position papers themselves are worthy of publication, and many who have participated in well-conducted multisite RCTs consider the production and review of such position papers a most valuable type of research education.

Finally, when even such position papers do not settle the issue, the study PI might consult with two or three of the most highly regarded experts on the specific topic or ask for input from the DSMB. Again, the outcome of such consultation would be included in the documentation of the study. Such effort (1) minimizes the chance of a serious error in the design or execution of the multisite RCT and (2) prevents interminable argument on the subject.

**Conclusion**

The multisite RCT is a research approach that is ideal in some situations and totally unwarranted in others. Even when the situation is ideal, a multisite RCT can be done well or badly. There is now a great deal of information on the circumstances under which a multisite RCT is warranted: where single-site RCTs produce heterogeneous and unconvincing results unlikely to lead to consensus. Then the only avenue to consensus may be the multisite RCT. There is also by now a long history of successful multisite RCTs providing the models for future RCTs, even in areas of research in which multisite RCTs have not to date been common. By the same token, there is now a long history of less than totally successful multisite RCTs and enough researchers with experience in those RCTs that many of the pitfalls should be discussed and avoided.

**References**


Harris, R.J. Significance tests have their place. Psychological Science, 8(1):8–11, 1997.

The Author

Helena Chmura Kraemer, Ph.D., is Professor of Biostatistics in Psychiatry, Stanford University, Stanford, CA.