Why Does the Randomized Clinical Trial Methodology So Often Mislead Clinical Decision Making? Focus on Moderators and Mediators of Treatment

Helena Chmura Kraemer, PhD; and Robert D. Gibbons, PhD

Psychiatric Annals, Volume 39, Issue 7, July 2009

CME EDUCATIONAL OBJECTIVES

1. Review moderators and mediators of treatment efficacy and effectiveness.
2. Identify limitations of randomized clinical trials for generalizing to the population.
3. Identify moderators and mediators for treatment-related effects.

ABOUT THE AUTHOR

Helena Chmura Kraemer, PhD, is Professor of Biostatistics in Psychiatry (Emerita), Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, California. Robert D. Gibbons, PhD, is with the Center for Health Statistics, University of Illinois at Chicago.

Address correspondence to: Helena Chmura Kraemer, PhD, Professor of Biostatistics in Psychiatry (Emerita), Department of Psychiatry and Behavioral Sciences, Stanford University, 1116 Forest Avenue, Palo Alto, CA 94301; or e-mail hckhome@pacbell.net.

Dr. Kraemer has disclosed no relevant financial relationships. Dr. Gibbons disclosed the following relevant financial relationships: National Institutes of Mental Health: research grant recipient (R56-MH078580 and R01-MH8012201).

doi: 10.3928/00485713-20090625-06

PARTICIPANT ATTESTATION

___ I certify that I have read the article(s) on which this activity is based, and claim credit commensurate with the extent of my participation.

COMMERCIAL BIAS EVALUATION

Please rate the degree to which the content presented in this activity was free from commercial bias.

	No bias				Significant bias
	5	4	3	2	1

Comments regarding commercial bias: ______________________________

____

INSTRUCTIONS

1. Review the stated learning objectives of the CME articles and determine if these objectives match your individual learning needs.

2. Read the articles carefully. Do not neglect the tables and other illustrative materials, as they have been selected to enhance your knowledge and understanding.

3. The following quiz questions have been designed to provide a useful link between the CME articles in the issue and your everyday practice. Read each question, choose the correct answer, and record your answer on the CME REGISTRATION FORM at the end of the quiz. Retain a copy of your answers so that they can be compared with the correct answers should you choose to request them.

4. Type your full name and address and your date of birth in the space provided on the CME REGISTRATION FORM.

5. Complete the evaluation portion of the CME REGISTRATION FORM. Forms and quizzes cannot be processed if the evaluation portion is incomplete. The evaluation portion of the CME REGISTRATION FORM will be separated from the quiz upon receipt at PSYCHIATRIC ANNALS. Your evaluation of this activity will in no way affect the scoring of your quiz.

6. Your answers will be graded, and you will be advised whether you have passed or failed. Unanswered questions will be considered incorrect. A score of at least 80% is required to pass. Your certificate will be mailed to you at the mailing address provided. Upon receiving your grade, you may request quiz answers. Contact our customer service department at (856) 994-9400.

7. Be sure to complete the CME REGISTRATION FORM on or before July 31, 2010. After that date, the quiz will close. Any CME REGISTRATION FORM received after the date listed will not be processed.

8. This activity is to be completed and submitted online only.

Indicate the total time spent on the activity (reading article and completing quiz). Forms and quizzes cannot be processed if this section is incomplete. All participants are required by the accreditation agency to attest to the time spent completing the activity.

CME ACCREDITATION

This CME activity is primarily targeted to patient-caring physicians specializing in psychiatry. There are no specific background requirements for participants taking this activity. Learning objectives are found at the beginning of each CME article.

This activity has been planned and implemented in accordance with the Essential Areas and policies of the Accreditation Council for Continuing Medical Education through the joint sponsorship of Vindico Medical Education and PSYCHIATRIC ANNALS. Vindico Medical Education is accredited by the ACCME to provide continuing medical education for physicians.

Vindico Medical Education designates this educational activity for a maximum of 3 *AMA PRA Category 1 Credits*™. Physicians should only claim credit commensurate with the extent of their participation in the activity.

FULL DISCLOSURE POLICY

In accordance with the Accreditation Council for Continuing Medical Education's Standards for Commercial Support, all CME providers are required to disclose to the activity audience the relevant financial relationships of the planners, teachers, and authors involved in the development of CME content. An individual has a **relevant financial relationship** if he or she has a financial relationship in any amount occurring in the last 12 months with a commercial interest whose products or services are discussed in the CME activity content over which the individual has control. Relationship information appears at the beginning of each CME-accredited article in this issue.

UNLABELED AND INVESTIGATIONAL USAGE

The audience is advised that this continuing medical education activity may contain references to unlabeled uses of FDA-approved products or to products not approved by the FDA for use in the United States. The faculty members have been made aware of their obligation to disclose such usage.

HOW TO OBTAIN CME CREDITS BY READING THIS ISSUE

This CME activity is primarily targeted to patient-caring physicians specializing in psychiatry. Physicians can receive *AMA PRA Category 1 Credits*™ by reading the CME articles in Psychiatric Annals and successfully completing the quiz at the end of the articles. Complete instructions are given subsequently. Educational objectives are found at the beginning of each CME article.

CME ACCREDITATION

This activity has been planned and implemented in accordance with the Essential Areas and policies of the Accreditation Council for Continuing Medical Education through the joint sponsorship of Vindico Medical Education and Psychiatric Annals. Vindico Medical Education is accredited by the ACCME to provide continuing medical education for physicians.

Vindico Medical Education designates this educational activity for a maximum of 3 *AMA PRA Category 1 Credits*™. Physicians should only claim credit commensurate with the extent of their participation in the activity.

FULL DISCLOSURE POLICY

In accordance with the Accreditation Council for Continuing Medical Education's Standards for Commercial Support, all CME providers are required to disclose to the activity audience the **relevant financial relationships** of the planners, teachers, and authors involved in the development of CME content. An individual has a relevant financial relationship if he or she has a financial relationship in any amount occurring in the last 12 months with a commercial interest whose products or services are discussed in the CME activity content over which the individual has control. Relationship information appears at the beginning of each CME-accredited article in this issue.

UNLABELED AND INVESTIGATIONAL USAGE

The audience is advised that this continuing medical education activity may contain references to unlabeled uses of FDA-approved products or to products not approved by the FDA for use in the United States. The faculty members have been made aware of their obligation to disclose such usage.

Financial Disclosures: Stanley Caroff, MD, has disclosed no relevant financial relationships. John M. Davis, MD, has disclosed no relevant financial relationships. Jan Fawcett, MD, has disclosed the following relevant financial relationships: Merck: Member of Editorial Advisory Board, Merck Manual. Paula Hensley, MD, has disclosed no relevant financial relationships. Andrew A. Nierenberg, MD, has disclosed the following relevant financial relationships: Abbott, Brain Cells Inc., Bristol Myers Squibb (BMS), Eli Lilly, GlaxoSmithKline (GSK), Innapharma, PGx, Ortho-McNiel Janssen, Novartis, Pfizer, Sepracor, Shire, and Somerset: Member of Advisory Board/Consultant: Massachusetts General Hospital (MGH) Structured Interview Guide to the Montgomery-Åsberg Rating Scale (MADRS), and the Clinical Positive Affect Scale, licensed exclusively to the MGH Clinical Trials Network and Institute. Copyright Holder: BMS, Cederroth, Cyberonics, Eli Lilly, Forest, GSK, Ortho-McNeil Janssen; Lichtwer, NARSAD, National Institute of Mental Health; Pfizer, Stanley Foundation, Wyeth: Research Grant Recipient; and BMS, Cyberonics, Eli Lilly, Forest, GSK, Wyeth: Member of Speakers' Bureau. The staff of *Psychiatric Annals* have disclosed no relevant financial relationships.

INSTRUCTIONS

1. Review the stated learning objectives of the CME articles and determine if these objectives match your individual learning needs.

2. Read the articles carefully. Do not neglect the tables and other illustrative materials, as they have been selected to enhance your knowledge and understanding.

3. The following quiz questions have been designed to provide a useful link between the CME articles in the issue and your everyday practice. Read each question, choose the correct answer, and record your answer on the CME REGISTRATION FORM at the end of the quiz. Retain a copy of your answers so that they can be compared with the correct answers should you choose to request them.

4. Type or print your full name and address and your date of birth in the space provided on the CME REGISTRATION FORM.

5. Complete the evaluation portion of the CME REGISTRATION FORM. Forms and quizzes cannot be processed if the evaluation portion is incomplete. The evaluation portion of the CME REGISTRATION FORM will be separated from the quiz upon receipt at PSYCHIATRIC ANNALS. Your evaluation of this activity will in no way affect the scoring of your quiz.

6. Send the completed form, with your $25 payment (check, money order, or credit card information) to: VINDICO MEDICAL EDUCATION, PO Box 36, Thorofare NJ 08086. Payment should be made in US dollars drawn on a US bank.

7. Your answers will be graded, and you will be advised whether you have passed or failed. Unanswered questions will be considered incorrect. A score of at least 80% is required to pass. Upon receiving your grade, you may request quiz answers. Contact our customer service department at (856) 994-9400.

8. Be sure to mail the CME REGISTRATION FORM on or before the deadline listed. After that date, the quiz will close. Any CME REGISTRATION FORM received after the date listed will not be processed.

Indicate the total time spent on the activity (reading article and completing quiz). Forms and quizzes cannot be processed if this section is incomplete. All participants are required by the accreditation agency to attest to the time spent completing the activity.

EDUCATIONAL OBJECTIVES OVERVIEW

If any readers of *Psychiatric Annals* read any psychiatric articles in any journal so that they can keep up to date and use the best-available evidence in their clinical practice, then the statistical papers in this issue are essential not only to read, but also to re-read, study, and master. What do these excellent statistical papers have to do with clinical psychiatric practice? They contain the tools that will allow you to critically appraise and interpret the psychiatric literature. They demystify the statistics used to generate the evidence

TABLE OF CONTENTS

RESPONSIBILITY FOR STATEMENTS

All opinions expressed by authors and quoted sources are their own and do not necessarily reflect the opinions of the editors, publishers, or editorial boards of *Psychiatric Annals* or its employees, Vindico Medical Education or its employees, or the University of New Mexico. The acceptance of advertising in no way implies endorsement by the editors, publishers, or editorial boards of *Psychiatric Annals* .

The material presented at or in any *Psychiatric Annals* or Vindico Medical Education continuing education activity does not necessarily reflect the views and opinions of Vindico Medical Education or *Psychiatric Annals*. Neither *Psychiatric Annals*, Vindico Medical Education, nor the faculty endorse or recommend any techniques, commercial products, or manufacturers. The faculty/authors may discuss the use of materials and/or products that have not yet been approved by the U.S. Food and Drug Administration. Articles are intended for informational purposes only and should not be used as the basis of patient treatment. All readers and continuing education participants should verify all information before treating patients or utilizing any product.

CME

Why Does the Randomized Clinical Trial Methodology So Often Mislead Clinical Decision Making? Focus on Moderators and Mediators of Treatment

Helena Chmura Kraemer, PhD; and Robert D. Gibbons, PhD

CME **EDUCATIONAL OBJECTIVES**

1. Review moderators and mediators of treatment efficacy and effectiveness.
2. Identify limitations of randomized clinical trials for generalizing to the population.
3. Identify moderators and mediators for treatment-related effects.

Helena Chmura Kraemer, PhD, is Professor of Biostatistics in Psychiatry (Emerita), Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, California. Robert D. Gibbons, PhD, is with the Center for Health Statistics, University of Illinois at Chicago.

Address correspondence to: Helena Chmura Kraemer, PhD, Professor of Biostatistics in Psychiatry (Emerita), Department of Psychiatry and Behavioral Sciences, Stanford University, 1116 Forest Avenue, Palo Alto, CA 94301; or e-mail hckhome@pacbell.net.

Dr. Kraemer has disclosed no relevant financial relationships. Dr. Gibbons disclosed the following relevant financial relationships: National Institutes of Mental Health: research grant recipient (R56-MH078580 and R01-MH8012201).

doi: 10.3928/00485713-20090625-06

To provide a foundation for our discussion on moderators and mediators of treatment efficacy and effectiveness, we begin with a detailed discussion of the strengths and weaknesses of randomized clinical trials (RCTs).

Clinician/patient: Why are so many treatments shown to have highly statistically significant effects in RCTs later shown to have so little benefit when used for my patients/me? What's wrong with RCT methodology?

IN DEFENSE OF RCTS

What leads to such false-positive results is not a problem with RCT methodology per se. To be blunt, many studies labeled as RCTs do not follow the "rules" of RCT methodology. Analysis is done, not on those randomized to treatment ("by intention to treat"), but on those who complete treatment, or adequately comply with treatment. Control of bias (eg, "blinding") is either absent or incomplete. When the original primary outcome is not statistically significant, researchers have been known to "cherry pick" among other available outcomes. All such tactics lead to false-positive results, indeed often misleading clinical decision-making; but the problem is not RCT methodology, but the lack

© 2009/Photos.com, JupiterImages Corporation

of compliance with RCT methodology. Consolidating Standards for Reporting Trials (CONSORT) guidelines,[1-4] and the current requirement to register RCTs before onset, recognize these problems and represent an effort to encourage compliance with RCT "rules" in order to provide clinicians and medical consumers greater assurance as to the credibility of their results.

However, among RCTs that are adequately designed and executed, misunderstanding of "statistical significance" continues to be a problem.[5-11] All that is usually shown by a statistically significant result is that some nonrandom difference was detected between the responses to the treatments compared, ie, T1 versus T2. That means that the design and sample size was large enough to detect a difference, not that the difference is of any clinical significance. To judge the clinical significance of a statistically significant finding, what is needed is an informative effect size. Although CONSORT guidelines for reporting RCTs recommend reporting an effect size with every *P* value, reporting effect sizes that clinicians and medical consumers can interpret is still not the norm.

For the purpose of this discussion, let us propose as such an effect size, "success rate difference" (SRD),[12,13] defined as the difference between the probabilities that a subject in the T1 group has a clinically preferable response to one in the T2 group (T1 > T2) and that a subject in the T2 group has a clinically preferable response to one in the T1 group (T1 < T2), symbolically:

SRD = Probability (T1 > T2) – Probability (T1 < T2)

This effect size is zero (SRD = 0) if patients are as likely to benefit from T1 as from T2. SRD = 1 if every single person given T1 has a clinically preferable response to every single person given T1; SRD = -1 if every single person given T2 has a clinically preferable response to every single person given T1.

In fact, the SRD in a RCT is never absolutely zero.[14-16] By the time there is rationale and justification to propose a RCT to compare T1 versus T2, the chance of an absolutely zero SRD is itself zero. At the other extreme, not even the best of known treatments have SRD equal either 1 or -1. As a general guideline, SRD = + 0.1 would be considered a "small" effect, SRD = + 0.3 a "medium" effect; and SRD = + 0.5 a "large" effect, but what size effects we, as medical consumers, should get excited about might differ depending on the severity of the condition to be treated, the consequences of inadequate treatment, the vulnerability of the population, and the costs or risks of treatment, etc. For example, for a safe vaccine preventing AIDS, SRD = .01 would be an earth-shaking result.

For a cold medication, SRD = .5 might not be of any great consequence.

There are other effect sizes: Number Needed to Treat (NNT), for example, is the number of subjects you'd expect to have to treat with T1 to have one more success than if you'd treated them with T2.[13,17-19] NNT is an effect size clinicians, policy makers, and medical consumers often prefer because it is expressed in terms of number of patients rather than probability points. But, as is true of most interpretable effect sizes, NNT is directly related to SRD: NNT equals 1/SRD. Thus "small," "medium," and "large" in NNT terms are 9, 4, and 2. On the other hand, the most commonly used effect size is Cohen's d,[20] the standardized mean difference between the two treatment group means, when outcomes have a normal distribution. Again, when d is appropriately used, d is directly related to SRD: SRD = 2 phi(d/ 2) -1, where phi() is the cumulative distribution of the standard normal distribution. Now "small," "medium," and "large" in d terms are 0.2, 0.5, 0.8.[20]

In short, part of the clinicians'/patients' problem with well-done RCTs is that little or no attention is paid to effect sizes and their clinical impact. Until it becomes common practice to follow CONSORT guidelines and report an interpretable effect size with every *P* value, and to discuss its magnitude in the appropriate medical context, misinterpretations of the clinical significance of RCT findings will continue.

Even among well-done RCTs that do report effect sizes, there are still problems. Many RCTs are done comparing a new active treatment (T1) with an inactive placebo (T2), even when clinicians in the community have access to old effective treatments. In such cases, T2 amounts to withholding treatment, which many researchers consider unethical.[21-24] Ethical or not, T1 that is significantly better than an inactive placebo may be significantly worse than whatever treatments are already being used in the community.

Moreover, many RCTs impose very limiting inclusion/exclusion criteria for participation in the study. RCTs have been known to exclude 90% of those with the condition to be treated in the RCTs, focusing on the 10% most likely to best respond to T1.[25] Clearly, the result of any such RCT is likely to exaggerate the effectiveness of T2 for the 90% excluded from the study.

Such studies may have a major impact on future treatments of the condition ...

These problems, relating to sample selection and choice of control treatment, have been well-recognized in recent years. In so-called "effectiveness" RCTs,[26] only those in the target population who will be harmed, or known from previous studies not to be helped by T1 or T2 ("clinical equipoise")[27] are excluded. There, the choice of control group in a RCT is now often "treatment as usual," or some standard of treatment. CONSORT guidelines now require that exclusions and the reasons for exclusion be reported. Where researchers exclude many more patients with the condition than they include, the results of the study should be taken with a grain of salt before applying to general clinical decision making.

Finally, RCTs are done for a variety of purposes. Those done by pharmaceutical companies for submission to the Food and Drug Administration (FDA) often use stringent inclusion/exclusion criteria and placebo controls, focusing solely on statistical significance, since that is what is required by the FDA for licensing. Such studies cannot establish the efficacy/effectiveness of a treatment for clinical decision making in the community. Moreover, scientists often do RCTs, not to influence clinical decision making, but to understand the pharmacokinetics of drugs, the etiology or expression of the condition to be treated, or to resolve other such basic science questions. Such studies may have a major impact on future treatments of the condition, but should not be used to change present clinical decision-making.

Therefore, the exaggeration of effectiveness that clinicians, medical policy makers, and patients perceive in RCTs is not a problem with RCT methodology per se, but either in uses of that methodology for purposes other than influencing medical decision making (which may well be right and proper), or misuses of the methodology (which are not). RCT methodology seldom leads to false-positive results when used appropriately.

Clinician/patient: So you are saying that RCT methodology is perfect? That if we simply focused on the effect sizes of well-done, well-reported, relevant RCTs, we would be assured of highly effective treatments?

THE LIMITATIONS OF RCTS

Yes and no. Indeed, focusing on such RCTs, particularly on those whose results have been independently replicated, would protect well against false-positive results. RCT methodology has developed over many years by identifying weaknesses in the existing methodology and correcting them, and thus is never likely to achieve perfection. It will continue to evolve. However, to date, most such concern has focused on avoiding false-positive results (ie, the exaggeration of effectiveness of treatments on which your initial question focused). The as-yet unsolved problems with well-done and relevant RCTs lie not with false positives but with false negatives. Currently, if clinicians, medical policy makers, and patients focused on well-done, well-reported, relevant RCTs in the research literature, what they would most likely find is that effect sizes, even those that are statistically significant, tend to be very small, so small that we might be led to believe that all our current treatments are ineffective.

This problem stems from the fact that an effect size comparing T1 versus T2 in a RCT is an average effect over all the individual patients in the population. It is seldom, if ever, true that "one size fits all," that the effect size in a population, particularly a heterogeneous one, applies to every individual patient within that population.

Generally within every population, there are some patients for whom T1 is clinically significantly better than T2 (T1 > T2), some for whom T1 and T2 are essentially clinically equivalent (T1 = T2), and some for whom T1 is clinically significantly worse than T1 (T1 < T2). Whichever of the two non-equivalent subpopulations (T1 > T2 or T1 < T2) is predominant in the population determines which of T1 and T2 will overall appear better in that population, an effect that will be detected as "statistically significant," provided the sample size is large enough. However, the existence of the other two subpopulations will always dilute the overall effect size. The information necessary to identify the subpopulations, thus to "target" T1 and T2 optimally, is contained in "moderators" of treatment effects, but moderators, until recently, have been largely ignored in RCTs.

Moreover, T1 and T2 in a RCT are defined by fixed protocols, again a "one size fits all" approach. Traditionally, this might mean a fixed daily dosage of a drug, or a fixed sequence of psychotherapy sessions, the same for all patients. More recently, the protocol might be a strategic plan with some planned flexibility for individual patients, meant to "tailor" the treatment to individual needs of patients in the population. The information needed to structure such a flexible strategy is contained in the "mediators" of treatment. However, mediators too, until recently, have enjoyed little attention in RCT methodology.

Accordingly, issues related to moderators and mediators in RCTs are the focus of the remainder of this discussion. The terms "moderator" and "mediator" have been around for at least 50 years, but used inconsistently and often idiosyncratically by researchers, largely outside medical research. Consequently their value and importance was not recognized. In 1986, Baron and Kenny wrote a seminal paper[28] proposing specific conceptual definitions for these terms and statistical procedures for their implementation. The conceptual definitions proved valuable, but the implementation proved confusing.[29,30] Around 2000, members of the MacArthur Network on Developmental Psychopathology, proposed modified implementation procedures based on the Baron and Kenny conceptual definitions.[29-33] The following discussion is based on that so-called MacArthur Model in the context of RCTs.

We will first present a general introduction to moderators and mediators, defining each and distinguishing between them. A brief consideration of the types of research studies that would be needed to detect (hypothesis-generating) and to establish (hypothesis-testing) moderators and mediators then follows. Such studies follow standard design and analysis principles, but are often more difficult, costly, and time consuming than the standard RCT. For this reason there is a reluctance to invest in such studies, both by funders, reviewers and researchers. Consequently, we focus subsequent discussion on the problems involved in continuing to ignore the issues of moderators and mediators of treatments, concluding with a discussion of some of the limitations inherent in our discussion.

MODERATORS AND MEDIATIONS: DEFINITIONS

"Moderation" and "mediation" refer to certain types of associations relating three random variables in the population, a target variable (T), an outcome variable (O), and a third variable (M) that might be a moderator of T on O (M = Mo), or a mediator of T on O (M = Me). In RCTs, the target variable T is the random choice of treatment (T1 versus T2) made for each participant, and the outcome variable O is the primary outcome of treatment proposed in the "a priori" hypothesis that determines which of the two treatments is clinically preferred. (A very common error is describing a variable as a "moderator" or "mediator" without specifying the target or outcome variable.)

Moderator

A moderator of T on O is a variable that helps to identify on whom or under what conditions T has an effect on O. To show that Mo moderates the effect of T on O, it must be shown that 1) Mo precedes T which, in turn, precedes O, 2) Mo and T are independent, and 3) if the population is stratified by Mo, the effect size of T on O is different in different strata. The first two criteria indicate that in a RCT any baseline (pre-randomization) variable is eligible to be a moderator of treatment choice on any outcome. All that must be shown is that criterion 3) is satisfied.

For example, gender moderates the effect of choice of treatment (T1 versus

T2), if the effect size of T on O is different for men than it is for women. Age moderates the effect of choice of treatment if the effect size of T on O is different among subpopulations, say 25 to 34 years, 35 to 44 years, 45 to 54 years, etc. Most interesting, a gene may moderate the effect of choice of treatment on O if the effect size of T on O differs between those with one genotype from those with another at that gene locus.[34]

Mediator

A mediator of treatment (T) on outcome (O) explains how or why T has an effect on O. To show that a variable mediates the effect of T on O, it must be shown that 1) T precedes Me, which, in turn, precedes O, 2) T and Me are associated, and 3) the effect of T on O can be explained wholly (total mediator) or in part (partial mediator) by Me. Thus in an RCT, any event or change that occurs after treatment assignment but before determination of the outcome is eligible to be a mediator. What needs to be shown is first, that Me is associated with T (and thus possibly caused by choice of treatment), and then that all or part of the effect size of T on O is explained by the differential effects of T on Me.

For example, compliance with treatment is often a mediator of the effect of treatment on outcome.[35] This occurs if patients are more likely to comply with one treatment than with the other, and that at least some of the differential effect of the two treatments is associated with this difference in compliance. If this were so, one might consider improving the preferred treatment by including in its protocol some component directed to improving compliance. Even more interesting would be a situation, for example, where two drug treatments (T1 and T2) for children with ADHD, predicted differential changes in parenting behavior (Me), and that the change in symptoms proposed as the outcome measure (O) associated with treatment, is explained, at least in part, by these differential changes in parenting behavior. What such a result would suggest is that the causal mechanism by which change in symptoms in the child is induced by a treatment may be enhanced by inducing a greater change in parenting behavior.

Because of RCT structure, it is reasonable to infer that any effect of T on M is causal, and that any effect of T on O is causal, but there is no reason to infer that any effect of M on O is causal. A common error is to use the term "mediator" as a synonym for "cause," which is not necessarily true. All factors in a causal chain leading from treatment to outcome are mediators, but not all mediators are factors in that causal chain. The importance of mediators for clinical decision-making lies in the suggestion they convey about possible causal factors in such a chain.

According to these definitions, the same factor cannot be both a moderator and mediator of T on O, because a moderator precedes and is independent of T and a mediator follows and is associated with T. Based on the Baron and Kenny definitions of moderator/mediator there has been much discussion of single factors that serve both roles, and has been one of the more obvious difficulties with their implementation. However, it is possible, for example, that a factor, such as social support at baseline, may moderate T on an outcome such as change in symptoms, and that change of social support during treatment may mediate T on the same outcome. However, social support at baseline and change in social support during treatment are distinct variables, and may not even be highly associated with each other.

MODERATORS AND MEDIATORS: DETECTION AND DOCUMENTATION

There are many different statistical methods that might be used to detect or to test for moderators and mediators,[33] but currently a linear model is still the most common and useful approach. For example in a Multiple Linear Regression model, with O a scaled outcome:

$$O = b_0 + b_1T + b_2M + b_3TM + e$$

where e is an error term independent of T and M, with mean zero and variance V.

Because there is an interaction term in the model (TM), how one codes T and M changes the meaning of all the coefficients other than b_3.[36] For moderator/mediator, an analysis in an RCT comparing T1 and T2, T is always coded +/- 1/2 for the two treatments. M is coded according to what is known at the time of randomization. Thus, for a moderator (a baseline variable), binary Mo is coded as +/- 1/2, and scaled Mo is coded as deviations from the overall mean or median of M. For a mediator (an event or change following onset of treatment), binary Me is coded 0/1, and ordinal Me as deviations from 0.

Then, the effect size of T on O measured as Cohen's d,[20] the usual effect size with a linear model, is:

$$d = \frac{b_1 + b_2\,(M_1 - M_2) + b_3\,(M_0 - c)}{\sqrt{V + .5[b_2 + b_3 / 2)^2\,W_1 - (b_2 - b_3/2)^2W_2]}}$$

where M_1 and M_2 are the means of M in the T1 and T2 groups, M_0 is the average of M_1 and M_2, W_1 and W2 the variances of M in those two groups, and c the centering value.

Moderators and Nonspecific Predictors

If M is independent of T, as is required for a moderator, $M_1 = M_2 = M_0 = c$, and $W_1 = W_2 = W$. The effect size comparing T1 versus T2 in the subpopulation with any particular value of M, M = m, is $d = (b_1 + b_3(m\text{-}c))/V^{1/2}$. Thus, the effect sizes of treatment in the subpopulation with M=m depends on what m is, provided b_3 is not equal to zero: one would have to demonstrate a significant interaction effect of T and Mo on O. In this case the overall effect size is $d/(V{+}b_2b_3W)^{1/2}$,which does not depend on any differential effects of T1 versus T2 on M.

If, in this situation with a baseline variable in a RCT, $b_3 = 0$ (hence M not a moderator of T on O), but b_2 is not equal to zero, M is then called a "nonspecific predictor of the effect of T on O."[30] In this situation, the overall effect size is $d/V^{1/2}$, which does not depend either on common or differential effects on M, and the effect size of T is the same for all subpopulations matched on M. However, within both the T1 and T2 groups, outcome is predicted by M. Moderators of T on O are relatively uncommon; nonspecific predictors of the effect of T on O are quite common. A common error is to confuse a predictor of O within the T1 or T2 group with a moderator of T on O.

Whether a statistically significant moderation is clinically significant or not would depend on how much of the population lay in a range of values of Mo with clinically significant T1 > T2, and how much in the range with clinically significant T1 < T2. It is quite possible that there is a statistically significant moderation, but almost the entire population has T1 > T2. It is also quite possible that there is no overall treatment effect, because half the population has T1 ~ T2 and the other half T1 < T2. Thus, once again, showing statistically significant moderation only begins the discussion of its clinical significance. Comparing the effect sizes in the moderated subgroups and their impact on clinical decision-making matters.

Mediator

On the other hand, if M is an event or change occurring after T, that is associated with T, then M1 is not necessarily equal to M_2, c = 0, but M_0 is not necessarily equal to c, and finally, W_1 is not necessarily equal to W_2. Then it can be seen from the effect size above that portions of the effect size of T on O are associated with the effects of T on M, provided either b_2 or b_3 is nonzero. Thus to show mediation, one would have to reject the null hypothesis that $b_2 = b_3 = 0$. If both b_2 and b_3 are zero, then $d = b_1/V^{1/2}$, in which case, M has no affect on O, either directly, or indirectly, via T.

Once again, the demonstration of a statistically significant mediator effect does not necessarily connote clinical significance, but the issue of clinical significance of a mediator is much more complex than that of a moderator. The clinical value of a mediator is in the suggestion it carries that modifying T1, where T1 > T2, to have a greater effect on M would improve the effect size of T1 versus T2 on O. But that is not necessarily so.

It is important to have strong theoretical rationale and empirical justification for a moderator before attempting to propose or design a study to test a moderator hypothesis.

For example, if compliance is shown to be a mediator of T on O, including a component in a modified T1, say $T1^*$, that produces even greater compliance, may or may not result in an effect size of $T1^*$ versus T2 on O larger than that of T1 versus T2. Thus, the clinical significance of a mediator depends on whether in future RCTs those mediator-suggested modifications in T1 improve the effect size of T1 relative to T2.

MODERATORS AND MEDIATORS: IMPLEMENTATION OF STUDIES

Once there is theoretical rationale and empirical justification for proposing a moderator or mediator of some choice of treatment on a specified outcome, it is easy enough to design a study to test that hypothesis using standard methods.

Testing a Moderator Hypothesis

If one proposed that gender moderates the effect of T (T1 versus T2) on a outcome O, for example, one would stratify the population by gender to have approximately equal numbers of males and females (particularly in the study of disorders such as schizophrenia where the majority of patients would be male, or depression where the majority would be female), and a total number of subjects for adequate power to detect any clinically significant interaction effect. Almost inevitably, such a study would be logistically more difficult, time-consuming, and costly than any standard RCT that ignored moderators. For this reason, it is important to have strong theoretical rationale and empirical justification for a moderator before attempting to propose or design a study to test a moderator hypothesis.

Testing a Mediator Hypothesis for Clinical Significance

Because a mediator is of clinical importance only if it can be used to improve the effectiveness of treatment, to demonstrate a clinically significant mediator, one might design a new RCT, in which patients are randomly assigned to T1, the original preferred treatment to T2, and to $T1^*$, which would be T1 augmented by components designed to change Me in the direction of improving O. The goal would be to show that $T1^* > T1$. Generally, the effect sizes indicating clinical significance between $T1^*$ and T1 would now be relatively small, and such a study, for adequate power, would require a much larger sample size than would a standard RCT comparing two different treatments. However, if T1 has a moderate effect size compared to T2, T1* may have a large effect size compared with T2. Once again, strong theoretical rationale and empirical justification is necessary both to proposing and to designing a study to test a mediator hypothesis.

Detecting Moderators and Mediators: Exploratory Data Analysis

Where does the empirical justification necessary to propose and design a study to test a mediator/moderator hypothesis come from? Such empirical justification comes primarily from exploratory (hypothesis-generating) studies, secondary data analyses, following earlier RCTs, combined with theory, clinical observations, and results from basic science studies. Such exploratory studies are difficult to get funded and even more difficult to publish, often derisively called "fishing expeditions" by reviewers. Because of this, it is common that the findings of an exploratory study are presented in the research literature as if they were hypothesis-testing. Because the purpose of exploratory studies is to generate hypotheses, not conclusions, such studies are very likely to proliferate false-positive results. Consequently, inappropriately presented, they can impede scientific progress and increase the appearance of nonreplicable results in the research literature. Appropriately done, they can foster scientific and clinical progress, generating more powerful hypotheses for future RCTs and better designed studies to test those hypotheses.

Ideally, such exploratory studies are done post hoc, as secondary analyses, on completed RCTs testing other hypotheses. The costs of such exploratory studies then are minimal, related simply to the time and costs of additional data analyses.

COSTS AND RISKS OF IGNORING MODERATORS AND MEDIATORS

Moderators

Researchers and reviewers, on some level, have long been aware of the problems of ignoring moderators of treatment in a RCT. It is, for example, quite common that reviewers will ask how researchers proposing a RCT will "control for" or "adjust for" gender, age, ethnicity, educational level, severity, previous medical history or treatment, various genes, various measures of brain structure or function, etc., even in absence of empirical evidence that any of these baseline variables are likely to impact the outcome. This is problematic, for either "controlling for" such factors by oversampling certain subgroups in the population, or "adjusting for" such factors by inclusion as independent variables in linear models without consideration of interactions (eg, Analysis of Covariance), when the focus of interest is the overall effect of treatment in the population, can result in loss of power for testing, and in loss of precision for estimating that overall effect size. When some of the factors to be "controlled for," or "adjusted for," are in fact moderators, but interactions are ignored, particularly when those factors are associated with each other (multicollinearity) or interact with each other in their effect on the outcome, either "controlling for" or "adjusting for" such factors may not only cost power and precision, but may bias the estimation of the overall effect size. Such loss of power and precision is often incurred while simultaneously incurring additional logistic difficulty and cost to the study.

It is almost inevitable that site is a nonspecific predictor of outcome.

Here are two examples. First, there is growing recognition of the importance of multisite RCTs, where the sample is stratified by site, and subjects within each site are randomly assigned to T1 and T2, with the same protocol followed at all sites. In many cases, multisite RCTs are the only way to generate adequate power for testing, and precision for estimating, the overall effect size. More importantly, in all cases, multi-site RCTs are the only way to test the generalizability of results over sites (ie, whether site moderates the effect of treatment on outcome).

It is almost inevitable that site is a nonspecific predictor of outcome,[37] and it sometimes happens, despite efforts to ensure fidelity to the same protocols at all sites, that site is a moderator of outcome[38] (ie, that the T1 versus T2 effect size may differ from one site to another). Not only are the samples satisfying inclusion/exclusion criteria at the different sites not necessarily from exactly the same population (eg, differing in ethnicity, socioeconomic status), but the staffs that deliver the treatments and assess the outcomes are different from one site to another. Consequently, in a multisite RCT, there is always rationale and justification to support the hypothesis that site moderates the effect of treatment on outcome. Nevertheless, many multi-site RCTs ignore site in their analyses, which may lead to false conclusions.[36,39]

The second problem deals with underrepresented populations (women, minority ethnic groups) in RCTs. In proposals for RCTs submitted to the National Institutes of Health (NIH), there is a requirement that women and minorities be included. Since conclusions from most RCTs are likely to be applied by clinicians to women and minorities, this makes both scientific and political sense. Moreover, any sample from the site population that does not include women and minorities in the proportion they exist in that site's population is necessarily a biased sample. If gender and/or ethnicity moderate the effects of treatment on outcome, the results of such a biased RCT are also likely to be biased. Thus, removing any barriers to the participation of women and minorities makes scientific and political sense.

However, when RCTs are done at sites where there are few African-Americans, or Hispanic-Americans, etc., researchers are encouraged to oversample the least represented ethnic subgroups in their site population or to "enrich" their sample by recruitment of minority groups from other sources. To do so biases the sample relative to the local population. In absence of any moderating effects of gender and ethnicity, this merely means extra effort and cost, for the estimate of the overall effect size will remain unbiased. However, if gender and/or ethnicity moderate the effects of treatment on outcome, the extra effort and cost may result in biased tests and estimates of the overall treatment effect, and the overall treatment effect, biased or not, is likely to mislead clinical decision making for the minority groups. It would be preferable to require representative samples, including minority groups to the extent they exist in the population sampled, in initial RCTs. Then secondary data analysis would be done to investigate the possibility that ethnicity or gender moderates treatment effects. If no such indication is found, that would raise no question. If such indication were found, subsequent RCTs would be need to be done at sites with adequate representation of those minorities, and would stratify by ethnicity/gender to have adequate sample sizes in each stratum to test the moderator effects.

Mediators

To propose protocols for T1 and T2 in a RCT requires both a theoretical basis and, more important, empirical justification. Unsupported theory alone is a poor basis for a hypothesized difference between any two treatments. However, it frequently happens that we can show that T1 > T2 in some population without knowing exactly how or why this happens (mediators). For example, it was known for many years that aspirin (and long before that, willow bark) reduced pain, but the fact that the mediator of aspirin in its effect on pain involved inhibition of hormone-producing enzymes became known only around 1970. The absence of focus on mediators of treatments does not, as does the absence of focus on moderators of treatment, mislead research and the effects of research on clinical decision making. Absence of focus on mediators of treatment misleads and slows the development of more effective treatments.

CONCLUSION

To summarize the situation: RCT methodology, to date, has focused on avoiding false-positive results, and, when the RCT "rules" are followed, RCTs are quite successful in doing so. However, the current problems with RCTs lie, not so much in false-positive results, but in false-negative results. Two major factors that, when ignored, can lead to false-negative results are moderators of treatment on outcome (baseline variables that identify subgroups with different effect sizes of treatment on outcome) and mediators of treatment on outcome (events or changes that occur during treatment that help explain the how or why of effect sizes).

In absence of a strong rationale for hypothesizing a moderator or mediator, the best RCT design is a large simple RCT that would focus on the overall effect size. Exploratory studies done on that RCT's dataset after the a priori hypotheses were tested and the results reported, would then be used first to explore the possibility of moderators, and then, within moderated subgroups if any, the possibility of mediators. If then, moderators and mediators with theoretical rationale and empirical justification are found, subsequent RCTs would focus on documenting and elaborating their use in clinical decision-making.

It is important to note that subgroup analysis, as has usually been conducted, is not the same as a moderator analysis. In exploratory subgroup analysis, what typically is done is to stratify the sample, say separating men and women, and to test the treatment effect separately in the two strata. Then if the treatment effect is statistically significant among men, but not among women, this is often taken to mean that T1 > T2 for men but T1 = T2 for women. However, a statistically significant treatment effect for men may not mean that T1 > T2, and a nonstatistically significant treatment effect for women may occur even if T1 > T2. In fact, it may be that the effect size for women is larger than that for men, but that the sample size for women was much smaller than that for men, leading to results such as these. It is important that the effect sizes be compared between the groups, not *P* values examined separately for each group.

Moreover, estimation error of the effect sizes must be taken into account. The effect sizes for men and women may appear quite different, but if one or the other or both are based on small sample sizes, their precision is poor and the appearance of a difference may be misleading. It is important to test whether effect sizes significantly differ (interaction in a linear model), and only if they differ significantly, to consider whether the differences are such that would lead to different clinical recommendations either for targeting (moderators), or tailoring (mediators) the treatments. To date, subgroup analyses are all too common, and moderator analyses done more rarely.

In this review we have focused on the comparison between two treatments. Yet RCTs often involve randomization to more than two treatments: T1, T2, T3, etc. However, the baseline factors that moderate/mediate the effect of T1 versus T2 on outcome may be different from the baseline factors that moderate/mediate the effect of T1 versus T3, or the effect of T2 versus T3 on the same outcome. Thus in RCTs moderators/mediators should always be investigated between pairs of treatments.

We have focused only on one analytic approach: the multiple linear regression model. There are other statistical approaches depending on the type of outcome (eg, multiple logistic analysis for binary outcomes, Cox Proportional Hazard Model for survival, etc.). There are also nonparametric approaches for the simpler situations, such as binary moderators/mediators, or those with a very limited number of categories (ordered or nonordered).[33] Nevertheless, the state of analytic methods to be used for moderator/mediator analyses is yet in its infancy.

We have also focused on a single primary outcome. When there are multiple outcomes in a RCT, it will often be found that different factors moderate/mediate the effect of treatment on different outcomes. Selecting one primary outcome has long been a recommendation for RCTs, a primary outcome on which the ultimate recommendation for T1 versus T2 would be based. When there are multiple primary outcomes, without adjustment for multiple testing, there is a proliferation of false positives. When there is appropriate adjustment for multiple testing, power is sacrificed, and, without an increase in sample size, there is a proliferation of false negatives. If the sample size is increased to compensate for the loss of power associated with adjustment for multiple testing, it frequently happens that for some outcomes T1 > T2 and other outcomes T1 < T2. Then in absence of attention to the relative clinical impacts of those multiple outcomes occurring in the same patient, and to the frequency of such co-occurrences, results of RCTs become uninterpretable for clinical decision-making. Thus, attention to moderators/mediators only adds further emphasis to the recommendation for a single primary outcome measure in a RCT, preferably one that integrates all the benefits or harms individual patients might experience that are clinically essential to the choice between using T1 or T2.

In short, RCT methodology still has a way to go to be completely satisfactory as a basis of clinical decision-making, and probably always will, but many of its problems have already been recognized and are being dealt with, and other problems, such as moderators/mediators are being recognized to be dealt with in future RCTs.

REFERENCES

1. Rennie D. How to report randomized controlled trials: the CONSORT statement. *JAMA.* 1996;276(8):649.
2. Begg C, Cho M, Eastwood S, et al. Improving the quality of reporting of randomized controlled trials: the CONSORT statement. *JAMA.* 1999;276(8):637-639.
3. Altman DG, Schulz KF, Hoher D, et al; CONSORT Group. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med.* 2001;134(8):663-694.
4. Piaggio G, Elbourne DR, Altman DG, Pocock SJ, Evans SJ. Reporting of Noninferiority and Equivalence Randomized Trials: an extension of the CONSORT Statement. *JAMA.* 2006;295(10):1152-1160.
5. Kline RB. *Beyond Significance Testing: Reforming Data Analysis Methods in Behavioral Research.* Washington, DC: American Psychological Association; 2005.
6. Nickerson RS. Null hypothesis significance testing: a review of an old and continuing controversy. *Psychol Methods.* 2000;5(2):241-301.
7. Wilkinson L. The Task Force on Statistical Inference. Statistical Methods in Psychology Journals: Guidelines and Explanations. *American Psychologist.* 1999;54:594-604.
8. Thompson B. Journal editorial policies regarding statistical significance tests: heat is to fire as P is to importance. *Educational Psychology Review.* 1999;11:157-169.
9. Krantz DH. The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association.* 1999;44(448):1372-1381.
10. Shrout PE. Should significance tests be banned? Introduction to a special section exploring the pros and cons. *Psychological Science.* 1997;8(1):1-2.
11. Hunter JE. Needed: a ban on the significance test. *Psychological Science.* 1997;8(1):3-7.
12. Hsu LM. Biases of success rate differences shown in binomial effect size displays. *Psychol Methods.* 2004;9(2):183-197.
13. Kraemer HC, Kupfer DJ. Size of treatment effects and their importance to clinical research and practice. *Biol Psychiatry.* 2006;59(11):990-996.
14. Jones LV, Tukey JW. A sensible formulation of the significance test. *Psychol Methods.* 2000;5(4):411-414.
15. Meehl PE. Theory testing in psychology and physics: a methodological paradox. *Philosophy of Science.* 1967;34:103-115.
16. Meehl PE. Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology.* 1978;46:806-834.
17. Wen L, Badgett R, Cornell J. Number needed to treat: A descriptor for weighing therapeutic options. *Am J Health Sys Pharm.* 2005;62(19):2031-2036.
18. Cook RJ, Sackett DL. The number needed to treat: a clinically useful measure of treatment effect. *BMJ.* 1995;310(6977):452-454.
19. Altman DG. Confidence intervals for the number needed to treat. *BMJ.* 1998;317(7168):1309-1312.
20. Cohen J. *Statistical Power Analysis for the Behavioral Sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
21. Miller FG. Placebo-controlled trials in psychiatric research: an ethical perspective. *Biol Psychiatry.* 2000;47(8):707-716.
22. Lavori PW. Placebo control groups in randomized treatment trials: a statistician's perspective. *Biol Psychiatry.* 2000;47(8):717-723.
23. Hill AB. Medical ethics and controlled trials. *BMJ.* 1963;1(5337):1043-1049.
24. Rothman KJ, Michels KB. The continuing unethical use of placebo controls. *N Engl J Med.* 1994;331(6):394-398.
25. Humphreys K, Weingardt KR, Horst D, Joshi

AA, Finney JW. Prevalence and predictors of research participant eligibility criteria in alcohol treatment outcome studies, 1970-98. *Society for the Study of Addiction.* 2005;100:1249-1257.

26. Hoagwood K, Hibbs E, Brent D, Jensen P. Introduction to the Special Section: Efficacy and Effectiveness in Studies of Child and Adolescent Psychotherapy. *J Consult Clin Psychol.* 1995;63(5):683-687.
27. Freedman B. Equipoise and the ethics of clinical research. *N Engl J Med.* 1987;317(3)141-145.
28. Baron RM, Kenny DA. The Moderator-Mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *J Pers Soc Psychol.* 1986;51(6):1173-1182.
29. Kraemer HC, Stice E, Kazdin A, Kupfer D. How do risk factors work together to produce an outcome? Mediators, moderators, independent, overlapping and proxy risk factors. *Am J Psychiatry.* 2001;158(6):848-856.
30. Kraemer HC, Wilson GT, Fairburn CG, Agras WS. Mediators and moderators of treatment effects in randomized clinical trials. *Arch Gen Psychiatry.* 2002;59(10):877-883.
31. Kraemer HC, Frank E, Kupfer DJ. Moderators of treatment outcomes: clinical, research, and policy importance. *JAMA.* 2006;296(10):1-4.
32. Kraemer HC, Kiernan M, Essex MJ, Kupfer DJ. How and why criteria defining moderators and mediators differ between the Baron & Kenny and MacArthur approaches. *Health Psychology.* 2008;27(2):S101-S108.
33. Kraemer HC. Toward non-parametric and clinically meaningful moderators and mediators. *Stat Med.* 2008;27(10):1679-1692.
34. Murphy GM, Hollander SB, Rodrigues HE, Kremer C, Schatzberg AF. Effects of the serotonic transporter gene promoter polymorphism on mirtazapine and paroxetine efficacy and adverse events in geriatric major depression. *Arch Gen Psychiatry.* 2004;61(11):1163-1169.
35. The MTA Cooperative Group. Moderators and mediators of treatment response for children with attention-deficit/hyperactivity disorder: the Multimodal Treatment Study of Children with Attention-deficit/hyperactivity Disorder. *Arch Gen Psychiatry.* 1999;56(12):1088-1096.
36. Kraemer HC, Blasey C. Centering in regression analysis: a strategy to prevent errors in statistical inference. *Int J Methods Psychiatr Res.* 2004;13(3):141-151.
37. The MTA Cooperative Group. A 14-month randomized clinical trial of treatment strategies for attention-deficit/hyperactivity disorder: Modal Treatment Study of Children with Attention-deficit/hyperactivity Disorder. *Arch Gen Psychiatry.* 1999;56(12):1073-1086.
38. Infant Health and Development Program. Enhancing the outcomes of low birth weight, premature infants: a multisite randomized trial. *JAMA.* 1990;263(22):3035-3042.
39. Kraemer HC. Pitfalls of multisite randomized clinical trials of efficacy and effectiveness. *Schizophr Bull.* 2000;26(3):535-543.