International Journal of Methods in Psychiatric Research
*Int. J. Methods Psychiatr. Res.* 23(4): 401–410 (2014)
Published online 18 June 2014 in Wiley Online Library
(wileyonlinelibrary.com) **DOI:** 10.1002/mpr.1445

# A mediator effect size in randomized clinical trials

HELENA CHMURA KRAEMER[1,2]

1 Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA, USA
2 Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA, USA

**Key words**
mediators, randomized clinical trials, effect sizes

**Correspondence**
Helena Chmura Kraemer, 1116 Forest Avenue, Palo Alto, CA 04301, USA.
Telephone (+1) 650 328-7564
Email: hckhome@pacbell.net

Received 26 August 2013;
revised 22 March 2014;
accepted 8 April 2014

**Abstract**

To understand the process by which a treatment (T) achieves an effect on outcome (O) and thus to improve the effect of T on O, it is vital to detect mediators, to compare the impact of different mediators, and to develop hypotheses about the causal factors (all mediators) linking T and O. An index is needed to facilitate interpretation of the potential clinical importance of a mediator (M) of choice of T on treatment O in randomized clinical trials (RCTs). Ideally such a mediator effect size should (1) be invariant under any rescaling of M and O consistent with the model used, and (2) reflect the difference between the overall observed effect of T on O and what the maximal effect of T on O could be were the association between T and M broken. A mediator effect size is derived first for the traditional linear model, and then more generally for any categorical (ordered or non-ordered) potential mediator. Issues such as the problem of multiple treatments, outcomes and mediators, and of causal inferences, and the correspondence between this approach and earlier ones, are discussed. Illustrations are given of the application of the approach. *Copyright © 2014 John Wiley & Sons, Ltd.*

## Introduction

In clinical research there is growing realization of the importance of the identification of both moderators and mediators in randomized clinical trials (RCTs) comparing two treatments, T1 versus T2. As defined in the MacArthur approach (hereafter the MA approach), moderators of treatment in a RCT are baseline variables that help identify patients in whom the effect size of T1 versus T2 are different (Kraemer *et al.*, 2001, 2002, 2006, 2008). Thus moderators are the means of achieving the objectives of personalized medicine (Garber and Tunis, 2009; Lesko, 2007; Jain, 2002), delivering the appropriate treatment to each individual patient. Mediators of treatment describe events or changes that occur during treatment before the outcome of treatment is determined that help identify how or why differential effects of T1 versus T2 occur. Thus mediators, while themselves not necessarily causal, help to locate possible causal factors that link treatment choice to outcome, to identify processes by which treatments work and may help in developing new, more cost-effective, or safer treatments for the patients. Both the targeting function of moderators of treatment choice and the tailoring function of mediators of treatment choice are crucial to identifying and fostering the best possible clinical decision-making (King *et al.*, 2008). The focus of the present discussion is not on moderators, but on mediators in RCTs. It is important, however, at least to mention moderators in this context, for if there are strong moderators

of treatment (T) choice on outcome (O), the mediators of T on O may differ in the subpopulations defined by the moderator (M).

A continuing difficulty in addressing these questions is the inconsistent definitions of the terms "moderator" and "mediator" in the research literature. In a 1986 seminal paper, Baron and Kenny (1986) presented conceptual definitions for these terms. It was later found that ambiguity remained, and refinements to the definitions were proposed (the MA approach) that will here be used (Kraemer *et al.*, 2008). The remaining difficulties lie in the fact that many use the term "mediator" as a soft synonym for "causal factor". It is important to understand that the MA approach does not assume causality, nor does it draw inferences of causality, from its application. Any inference of a causal role for a mediator identified using the MA approach requires proof beyond that in the MA approach.

Clearly in medical research, what everyone wants to know is to what extent a treatment causes an outcome, and what the elements are of any causal chain linking treatment to outcome. Arguments have been ongoing for two millennia dating back to Aristotle as to how to prove causality. Current approaches to causality range from theoretical and philosophical work (e.g. Pearl, 2013) to a number of data analytic models (e.g. Rubin, 2004; MacKinnon, 2008). Carefully designed observational studies and complex modeling can certainly bring greater credibility to claims of causality but such inferences are always conditional on the assumptions made. In medical research, the "gold standard" to infer the causal effect of a treatment is not an analytic approach, but the RCT methodology. In particular, experience in medical research has shown that drawing inferences of causality for a mediator can be erroneous and cost the well-being and possibly the lives of patients (Silverman, 1998; Connolly, 2006).

The discussion of mediators here lies within the context of RCTs. Since patients are randomly assigned to treatment, it might be appropriate to infer that T causes O, and within the MA approach definition, that T causes M, but whether M causes O is always in question. There may always be an unrecognized moderator or an unrecognized proxy to M (a "confounder") that explains away any causal relationship between M and O. Thus in the MA approach, every element of a causal chain linking T to O is a mediator of T on O, but not every mediator of T on O is a link in such a causal chain. Discovery of mediators in the MA approach is therefore a precursor to seeking causal paths linking T with O.

Thus, it must be here emphasized that the issue here is not causality but mediation, and, where appropriate, the contrasts between various definitions of mediators will be briefly addressed in the present development so that the choices between such approaches might be clearer.

## Background

In the MA approach, to show that M mediates the effect of T on O in the MA approach, one must demonstrate three criteria in the population of interest:

- Temporal Precedence: T precedes M which precedes O (M is an event or change that occurs after treatment onset but before treatment outcome O is determined);
- Correlation: M and T are not independent (i.e. the distribution of M differs between the two treatment groups);
- Analytic: The effect size of T on O depends wholly or in part on the effect of T on M.

MacKinnon (2008) correctly points out that the MA approach essentially prohibits using cross-sectional data to study mediation, commenting "This would seem to prohibit the work of detectives, psychotherapists, and physicians who seek to untangle the process of events after they have occurred" (p. 71). However, establishing the time-line of events that have occurred is an essential part of the work of detectives, psychotherapists, and physicians.

The second and third criteria are common to most mediation approaches, but many approaches restrict the demonstration of the Analytic criterion to a linear model that assumes that the effect size of T on O conditional on M is the same for all values of M (no interaction in the linear model (e.g. MacKinnon, 2008, pp. 48–49). While a linear model is often useful (and will here be the initial approach), mediators are not always well-described by linear models, and if a linear model is appropriate, the interaction term may not be zero. Thus the non-parametric approach is likely to be more universally applicable.

An important lack in all mediator approaches to date is a definition of the effect size of the mediator. In recent years, the overuse and misuse of *p*-values has been increasingly recognized, and it is recommended that each *p*-value be accompanied by an estimated interpretable effect size and its confidence interval (Wilkinson, 1999). There have been suggestions for such an effect size (MacKinnon, 2008; Preacher and Kelley, 2011). To date, however, most proposed effect size measures depend on restrictive linear assumptions producing results that may be easy to interpret when the assumptions hold, but difficult otherwise.

Preacher and Kelley (2011) in considering the desiderata for good mediator effect sizes suggested that all effect sizes should be scaled appropriately, but what is considered "appropriate scaling" might differ among researchers.

However, if one can change the effect size simply by rescaling either the outcome or the mediator variable, the effect size becomes uninterpretable. It should be noted that any effect size is likely to change from one population to another; any effect size is population specific. The issue here is concern that within a single population, the effect size can be manipulated by changing the scaling of the outcome or mediator. Moreover, it becomes difficult, if not impossible, to compare multiple mediators measured on different scales of the same T on O in the same population. Thus we suggest that in a RCT it is required that:

(a) The mediator effect size be invariant under any rescaling of either the O or the proposed M that preserves the assumptions of the appropriate model.
(b) It equals the difference between the overall treatment effect size of T on O and the effect size if the connection between T and M were somehow severed.

As to what the treatment effect size might be in a RCT, either AUC (area under the receiver operating characteristic [ROC] curve comparing T1 and T2 responses) (Kraemer and Kupfer, 2006; Acion *et al.*, 2006; Grissom, 1994; McGraw and Wong, 1992) or SRD (success rate difference) (Kraemer and Kupfer, 2006; Hsu, 2004) is recommended because they are invariant under all monotonic transformations of O:

$$\mathrm{AUC} = \mathrm{Prob}(\mathrm{T1} > \mathrm{T2}) + 0.5\mathrm{Prob}(\mathrm{T1} = \mathrm{T2}) \quad (1)$$

$$\begin{aligned}\mathrm{SRD} &= \mathrm{Prob}(\mathrm{T1} > \mathrm{T2}) - \mathrm{Prob}(\mathrm{T2} > \mathrm{T1}) \\ &= 2\mathrm{AUC}–1\end{aligned} \quad (2)$$

where "T1 > T2" means that of two patients, one sampled from the T1 population and one sampled from the T2 population, the one from the T1 population has a clinically preferable response ("T1 = T2" means clinical equivalence of outcomes). AUC and SRD are essentially equivalent effect sizes (SRD = 2AUC − 1) defined for any outcome measure in a RCT, no matter how measured or how distributed, provided only that one can compare the individual outcomes of two patients, a necessary condition for a RCT outcome measure. AUC ranges from zero to one with null value ½, while SRD ranges from −1 (when every patient treated with T2 has a clinically preferable response to every patient treated with T1) to +1 (the reverse) with null value zero. SRD has a more easily interpretable scale and will here be used.

In a RCT with $N_1$ patients randomly assigned to T1 and $N_2$ to T2, one can always estimate AUC or SRD by making all $N_1N_2$ pairwise comparisons and estimating the probabilities indicated in Equations 1 and 2. If O is binary (success versus failure), SRD is the difference in the success rates of the two treatment groups. If O is ordinal, one can use the Mann–Whitney U-statistic, for AUC = $U/(N_1N_2)$ and SRD = 2AUC − 1. Finally, if O is normally distributed in the two treatment groups, then, Cohen's *d*, here defined as the difference in the two means divided by the square root of the average variance in the two groups (Kraemer and Kupfer, 2006), can be used:

$$\mathrm{SRD} = 2\Phi(d/\sqrt{2}) - 1, \quad (3)$$

where $\Phi()$ is the cumulative standard normal distribution function. This form of *d* is used because it relates directly to SRD whether the variances in the two groups are the same or not.

However, it should also be noted that if O is normally distributed with equal variances in the two treatment groups, *d* is equivalent to that form of d that uses the pooled standard deviation, or that form that used the standard deviation in the control/comparison group in the denominator. However, if the variances are not equal, then a pooled standard deviation estimates the standard deviation in neither treatment population, but a weighted average of the two that depends on the two sample sizes. If the sample sizes are equal, the pooled standard deviation would equal the average standard deviation. If the variances in the two treatment populations are not equal, using the standard deviation of the control group can be very misleading in indicating the degree of overlap between the treatment and control populations.

Preacher and Kelley (2011) correctly emphasize that an effect size is a population parameter estimated in a sample with a certain accuracy indicated by a confidence interval. Whether the population parameter suggested as an effect size conveys clinically or practically important information is the primary consideration and is here the focus. Testing the null hypothesis that the mediator effect size, however defined, equals zero is not the important issue. Once the appropriate effect size is defined, obtaining a tight confidence interval that may or may not include the null value is more crucial.

Preacher and Kelley (2011) focus on the linear model, assuming absence of interaction. Here, we first reconsider a linear model, but including the interaction term. Then we consider a non-parametric model if the mediator is categorical (ordered or unordered), with no assumptions about the distribution of the outcome measure O or the nature of the M versus O association in either T1 or T2, and demonstrate its application in special cases. In

conclusion, we discuss the implications of these results both in the RCT context.

## The linear model

Assume that M is a measure of change after onset of treatment with T1 or T2, and preceding determination of O, that M and O each have a normal distribution within both T1 and T2, possibly with different means and variances, and that:

$$M = a_0 + a_1 T + e^*,$$
$$O = b_0 + b_1 T + b_2 M + b_3 TM + e.$$

Here, conditional on T and M, e and e* are assumed to have normal distributions, the same for all T and M. The variance of e in either treatment group is V.

For clarity in interpreting the coefficients, T is coded +1/2 for T1 and −1/2 for T2 (Kraemer and Blasey, 2004). Absence of change corresponds to M = 0. So that the effect size is invariant under linear transformation of M (while ensuring the linearity model) M is recoded as $M/\sigma_A$, where $\sigma_A^2 = (\sigma_1^2 + \sigma_2^2)/2$, the average of the variances of M in the T1 and T2 groups and $P_1 = \sigma_1^2/\sigma_A^2$ and $P_2 = \sigma_2^2/\sigma_A^2$. Thus $P_1 + P_2 = 2$, and if the two M-variances are equal $P_1 = P_2 = 1$.

The expected value of O in the T1 and T2 groups are respectively:

$$b_0 + 0.5b_1 + b_2 M_1 + 0.5b_3 M_1,$$
$$b_0 - 0.5b_1 + b_2 M_2 - 0.5b_3 M_2,$$

where $M_1$ and $M_2$ are the means of recoded M in the two groups, and the mean difference is:

$$b_1 + b_2 \Delta M + b_3 M^*,$$

where $\Delta M = M_1 - M_2$ (the difference between the recoded M means), and $M^* = (M_1 + M_2)/2$, the average of the two recoded M means.

The within group variance in the T1 and T2 groups are respectively:

$$(b_2 + 0.5b_3)^2 P_1 + V, \text{ and}$$
$$(b_2 - 0.5b_3)^2 P_2 + V.$$

Then by the definition of Cohen's *d* as the mean difference divided by the square root of the average variance, and the fact that $SRD = 2\Phi(d/\sqrt{2}) - 1$, the SRD comparing the two treatments is given by:

$$\text{Overall SRD} = 2\Phi\left[\frac{b_1 + b_2\,\Delta M + b_3\,M^*}{\sqrt{\left(\left[(b_2 + 0.5b_3)^2 P_1 + V\right] + \left[(b_2 - 0.5b_3)^2 P_2 + V\right]\right)}}\right] - 1, \quad (4)$$

If M and T were independent, then $\Delta M = 0$, $M^*$ is the common mean of M in the two treatment groups, and the two variances of M are equal, i.e. $P_1 = P_2 = 1$. In this case,

$$\text{Null SRD} = 2\Phi\left[\frac{b_1 + b_3 M^*}{\sqrt{2\left(\left(b_2^2 + 0.25b_3^2\right) + V\right)}}\right] - 1. \quad (5)$$

The mediator effect size is the difference between these two:

$$\text{MedES} = 2\Phi\left[\frac{b_1 + b_2\,\Delta M + b_3\,M^*}{\sqrt{\left(\left[(b_2 + 0.5b_3)^2 P_1 + V\right] + \left[(b_2 - 0.5b_3)^2 P_2 + V\right]\right)}}\right] - 2\Phi\left[\frac{b_1 + b_3 M^*}{\sqrt{2\left(\left(b_2^2 + 0.25b_3^2\right) + V\right)}}\right]. \quad (6)$$

Overall SRD is the combination of the direct and indirect (via M) effect of T on O, Null SRD is the direct effect and MedES is the indirect effect (via M) on O.

In this model, it should be noted that, whether or not M is correlated with T, if $b_2 = b_3 = 0$, then both the Overall SRD and the Null SRD would equal $2\Phi[b_1/\sqrt{2V}] - 1$, and the MedES = 0. Thus if either the Correlation criterion or the Analytic criterion defining the mediator in the MA approach does not hold, the mediator effect size is zero, as appropriate.

One of the contentious differences between approaches to mediator analyses, when using the linear model, has been whether, if the linear model holds, mediation can


Int. J. Methods Psychiatr. Res. 23(4): 401–410 (2014). DOI: 10.1002/mpr

exist if $b_2 = 0$ when $b_3 \neq 0$. Some approaches avoid this argument by assuming that $b_3 = 0$. The MA approach addresses it because, as can be seen in Equation 4, the Overall effect size is influenced by non-zero $b_3$ even when $b_2 = 0$. However, if $b_2 = 0$, the MedES in a RCT is zero even for $b_3$ unequal to zero (Equation 6). Thus the MA approach is correct that $b_3$ does play a role in determining the effect size of treatment on O even when $b_2 = 0$, but when this happens, the mediator effect size is unlikely to have clinical importance. However, $b_2$ that appears in Equations 6–8, is that when the interaction term is included in model-fitting. If $b_3$ is non-zero in the population but omitted in model-fitting, the estimate of $b_2$ will usually be biased. How much this affects the mediator effect size depends both on the size of the interaction effect in the linear model and on the strength of the correlation between T and M. Thus this is not an argument for setting $b_3$ to zero in a linear model.

The estimation of the Mediator effect size so defined would proceed as follows:

- Step 1: A RCT in which $N_1$ patients are randomly assigned to T1 and $N_2$ patients to T2 is done, with M and O measured for each subject, "blinded" to T and to each other.
- Step 2: The overall effect size comparing T1 versus T2 is computed, ignoring M.
- Step 3: M is rescaled by dividing by the estimated $\sigma_A$, where $\sigma_A^2 = (\sigma_1^2 + \sigma_2^2)/2$, the average of the estimated variances of M in the two treatment groups. The difference in the means of rescaled M ($\Delta$M), and their average (M*) is computed. A linear regression is performed with O as the dependent variable and T (coded +1/2 for T1 and −1/2 for T2), M (rescaled) and the interaction T × M as the independent variables to obtain estimates of $b_1$, $b_2$, $b_3$, and the residual variance V.

Then

$$\text{Null SRD} = 2\Phi\left[\frac{b_1 + b_3 M^*}{\sqrt{2\left(\left(b_2^2 + 0.25 b_3^2\right) + V\right)}}\right] - 1.$$

- Step 4: The mediator effect size, MedES is the difference between the Overall AUC computed in Step 2 and the Null AUC computed in Step 3.
- Step 5: Bootstrap methods (Efron and Tibshirani, 1995; Efron and Gong, 1983) can be used to obtain a 95% two-tailed confidence interval for the mediator effect size. If the value zero is not inside that confidence interval, the mediator sample size is statistically significantly different from zero on a 5% two-tailed test.

It is to be noted that only in Step 3 are the linearity assumptions actually used, that is, they are used only in determining what the effect size would have been had T and M not been correlated.

## Illustration: parametric model

Goldin *et al.* (2012) report a RCT showing that change in positive self-views (M) mediate the effect of Cognitive-Behavioral Therapy (CBT) as compared to a Waiting List (WL) control for Social Anxiety Symptom Severity (O) for patients with Social Anxiety Disorder. Full description and data are there presented.

In that study, the 25th and 75th percentiles of M for the CBT group were 9.5 and 25, and those for the WL group were −2 and 9, clearly indicating the strong correlation between T and M, since the M-values of the two groups only overlap in the lower quartile of one and the upper quartile of the other. The within-group variances of M were not very different with $P_1 = 0.970$ and $P_2 = 1.030$. Finally the within group slopes (O on rescaled M) were only slightly and not statistically significantly different.

The mean O for CBT was 54.95, and for WL, 65.27. Cohen's *d* comparing the groups was 0.49, favoring CBT, and the Overall SRD equal to 0.27. The sample Null SRD then was found to be −0.06, and thus the sample MedES = 0.33.

A graphic display of the results appears in Figure 1. There it can be seen that the overall SRD strongly favors CBT over WL, indicated by the difference in length of the two vertical lines at M1 and M2. However, the Null SRD, indicated by the separation between the regression lines at M*, the effect size that would pertain if the correlation between T and M were broken, slightly favors WL over CBT. Thus here the mediator effect size is larger than the overall effect size, suggesting total mediation.

## The non-parametric model: categorical M, any O

There are many situations when the linear model assumptions will not hold. Suppose instead that M, an event or change occurring after onset of T and before O is determined, has C categories, coded $i = 1, 2, \ldots, C$. M may be binary (e.g. whether or not the patient fully complied with the treatment assigned), unordered categories (e.g. three types of quality of patient–doctor interaction during treatment), or ordered categories (e.g. a 3, 4, 5, … point ordinal scale, or a grouping of a continuous M into C ordered categories). Suppose that in T1 the probability that $M = i$ is $Q_i + \delta_i$, and in T2 is $Q_i - \delta_i$, for $i = 1, 2, \ldots, C$.

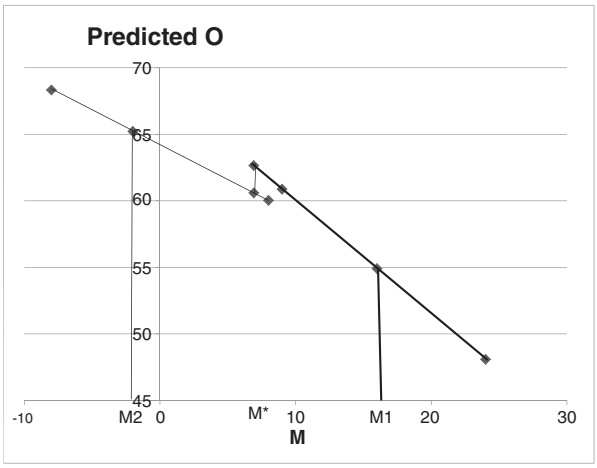


**Figure 1.** Graphical display of results in the Goldin *et al.* (2012) study. The regression lines extend through the 25th, 50th, 75th percentiles of M in each group (CBT, heavy line; WL, light line), and through M*. The Overall SRD is indicated by the difference in the lengths of the two vertical lines (at M1 and M2), and the Null SRD by the distance between the two regression lines at M*. The MedES is the difference between these two.

All $Q_i$ and $\delta_i$ are between zero and one, $\sum Q_i = 1$, $\sum \delta_i = 0$. Thus T and M are independent if and only if $\delta_i = 0$ for all $i$.

Then the overall effect size comparing responses in T1 versus T2 would be (Kraemer, 2008):

$$\text{Overall SRD} = \sum_{i,j}(Q_i + \delta_i)\big(Q_j - \delta_j\big)\text{SRD}(i,j), \quad (7)$$

where SRD($i$,$j$) is the effect size comparing O between patients with T1 in the subgroup with M = $i$ and those with T2 in the subgroup with M = $j$. Simplifying and separating terms that do not involve $\delta_i$ values from those that do:

$$\text{Overall SRD} = \sum_{i,j} Q_i Q_j \text{SRD}(i,j) + \sum_{i,j}\big(\delta_i Q_j - \delta_j Q_i - \delta_i\delta_j\big) \times \text{SRD}(i,j). \quad (8)$$

Thus the mediator effect size is:

$$\text{MedES} = \sum_{i,j}\big(\delta_i Q_j - \delta_j Q_i - \delta_i\delta_j\big)\text{SRD}(i,j). \quad (9)$$

MedES is invariant over any monotonic transformation of O and any permutation of the C categories of M, or any relabeling of those categories. SRD can be computed whether O is binary, categorical, dimensional, even multivariate, provided only that one can compare pairs of patients and decide which, if either, has a clinically preferable response. There are no assumptions about the distribution of M or O, and no assumptions that the association between M and O is monotonic, much less linear.

## Special Case 1: Binary M, Binary O

If the outcome itself is binary (success/failure), and M is also binary (C = 2), the situation is described in Table 1. Because O is binary, normality assumptions cannot hold. Now $Q_1 = Q$, $Q_2 = Q' = 1 - Q$, $\delta_1 = \delta$, $\delta_2 = -\delta$. The computations of the SRD($i$,$j$) are indicated in Table 1.

In this case, the Overall SRD is the difference in the success rates in the T1 andT2 groups, i.e.

$$\begin{aligned}\text{Overall SRD} = (Q + \delta)P_{1,1} + (Q' - \delta)P_{2,1} - (Q - \delta)P_{1,2} \\ -(Q' + \delta)P_{2,2} = [QP_{1,1} + Q'P_{2,1} - QP_{1,2} \\ -Q'P_{2,2}] + \delta[P_{1,1} - P_{2,1} + P_{1,2} - P_{2,2}],\end{aligned} \quad (10)$$

Thus the mediator effect size is:

$$\text{MedES} = \delta\big[P_{1,1} - P_{2,1} + P_{1,2} - P_{2,2}\big]. \quad (11)$$

Steps 1, 2, 4, and 5 described earlier for the linear model remain the same. To replace Step 3, Q is the estimated by the average of the T1 and T2 proportions with M = 1, and $P_{i,j}$ can be estimated by the proportions of success in each of the four cells to estimate the Null SRD, the first term in Equation 10. What is computed in Step 4 is Equation 11.

It is of interest to note that, if Equation 2 defining the linear model were here applied to the binary M and the $P_{i,j}$, $\delta$

**Table 1.** Binary M, Binary O. The probabilities of success (O) are $P_{i,j}$ and the probability that M = $i$ in T1 and T2 are shown in parentheses in each cell

| | T1 | T2 | |
|---|---|---|---|
| M = 1 | $P_{1,1}$ $(Q + \delta)$ | $P_{1,2}$ $(Q - \delta)$ | SRD(1,1) = $P_{1,1} - P_{1,2}$ |
| M = 2 | $P_{2,1}$ $(Q' - \delta)$ | $P_{2,2}$ $(Q' + \delta)$ | SRD(2,2) = $P_{2,1} - P_{2,2}$ |
| | SRD(1,2) = $P_{1,1} - P_{2,2}$ | SRD(2,1) = $P_{2,1} - P_{1,2}$ | |

Note: mediator effect size (MedES) = $\delta[P_{1,1} - P_{2,1} + P_{1,2} - P_{2,2}]$.


Int. J. Methods Psychiatr. Res. 23(4): 401–410 (2014). DOI: 10.1002/mpr

would correspond to $a_1$ and $[P_{1,1} - P_{2,1} + P_{1,2} - P_{2,2}]$ to $b_2$ in the linear model and MedES = $a_1b_2$, even though crucial assumptions of the linear model (normal distributions, equal variances) do not hold. Testing whether $a_1b_2 = 0$ is frequently the basis of testing for mediation in the MA approach (MacKinnon, 2008; MacKinnon *et al.*, 2002). However, this is a situation when users often switch to a Logistic Regression model rather than applying the Linear Regression model to the $P_{i,j}$, in which case MedES (Equation 11) will no longer correspond to $a_1b_2$ resulting from fitting that model.

### Special Case 2: Binary M, O satisfying the assumptions of the linear model

Suppose that the only deviation from linear assumptions is that M is binary (here coded 1/0 for M = 1 and M = 2). The probabilities that M = 1, 2 and the cell means in Table 2. The variance in each cell is V. Once again, Steps 1, 2, 4, and 5 remain the same. It is only Step 3 that differs.

Here, since the four cell means are described in terms of four parameters, the parameters can be directly estimated from the observed means, as shown in Table 2. V is estimated by the pooled within cell variance. The Null SRD is now (where computations of SRDs are shown in Table 2):

$$Q^2\mathrm{SRD}(1,1) + QQ'\mathrm{SRD}(1,2) + QQ'\mathrm{SRD}(2,1) + Q'^2\mathrm{SRD}(2,2), \quad (12)$$

to be used in Step 4 to estimate the mediator effect size. Step 5 remains the same. Now, even when a linear model holds, the MedES is *not* simply a function of $a_1b_2$.

It is also to be noted that Equation 12 is not the same as Equation 8, even though here both are based on a linear model. The difference lies in the normality assumptions for M. Under the assumptions underlying Equation 8, because M has a normal distribution in both groups, O too has a normal distribution within both T1 and T2. Depending on the distribution of M, however, the distribution of O in each group is a mixture of normal distributions, which affects the mediator effect size. In considering mediation, the distribution of M is as crucial as is that of O.

## Multiple treatments, multiple outcomes, multiple mediators

### Multiple treatments

The discussion thus far has focused on a RCT comparing two treatments, T1 versus T2. Many RCTs include more than two treatments, T1, T2, T3, …, and a reasonable question is how mediator analysis would be done under these circumstances.

What mediates the effect of treatment comparing T1 and T2 may be very different from what mediates the effect of treatment comparing T1 and T3. It may be that T1 and T2 have differential effects on M, while T1 and T3 do not, or that the direction or size of the mediation effect differs. Finally, the purpose of discovering strong mediators of treatment in a RCT is to provide insight as to how the protocol for the preferred treatment might be improved to achieve greater advantage of the preferred over the non-preferred treatment. The tactics available for consideration to improve the protocol for T1, T2, or T3 may differ even if the same mediator is identified for all pairwise comparisons.

**Table 2.** Binary M. Observations within each cell are normally distributed with the means indicated in each cell, according to a linear model, common within cell variance equal to V. The probability distribution of M in T1 and T2 are indicated in parentheses

| | T1 | T2 | |
|---|---|---|---|
| M = 1 | $M_{1,1} = b_0 + 0.5b_1 + b_2 + 0.5b_3$ ($Q + \delta$) | $M_{1,2} = b_0 - 0.5b_1 + b_2 - 0.5b_3$ ($Q - \delta$) | $\mathrm{SRD}(1,1) = 2\Phi((b_1 + b_3)/\sqrt{2V}) - 1$ |
| M = 2 | $M_{2,1} = b_0 + 0.5b_1$ ($Q' - \delta$) | $M_{2,2} = b_0 - 0.5b_1$ ($Q' + \delta$) | $\mathrm{SRD}(2,2) = 2\Phi((b_1)/\sqrt{2V}) - 1$ |
| | $\mathrm{SRD}(1,2) = 2\Phi((b_1 + b_2 + 0.5b_3)/\sqrt{2V}) - 1$ | $\mathrm{SRD}(2,1) = 2\Phi((b_1 - b_2 + 0.5b_3)/\sqrt{2V}) - 1$ | |

Estimation of parameters:
$b_1 = M_{2,1} - M_{2,2}$
$b_2 = (M_{1,1} - M_{2,1} + M_{1,2} - M_{2,2})/2$
$b_3 = (M_{1,1} - M_{2,1} - M_{1,2} + M_{2,2})$

In short, mediation should be considered separately for each pair of treatments, even when there are more than two treatments in a single RCT.

### Multiple outcomes

A major problem in RCTs in general is that of multiple outcomes. To protect against proliferation of false positives, some adjustment of *p*-values is usually recommended. However such adjustment is made, it costs power, necessitating larger sample size, thus usually increased cost and time. Even then, conflicting conclusions may result. It may be that T1 > T2 for some outcomes and T1 < T2 for others, confusing clinical decision-making.

Similarly, what mediates the effect of T1 versus T2 for one outcome (e.g. reduction of symptoms) may be completely different from what mediates the effect of T1 versus T2 for another (e.g. avoidance of serious side effects). A tactic that may improve the outcome for T1 over T2 for one outcome may actually deteriorate the outcome for T1 over T2 for another. The clinical impact of finding mediators of treatment in RCTs, consequently, will be diluted with multiple outcomes. Mediator analysis in RCTs is best done with a single outcome sensitive to clinically important differences among patient responses (Kraemer *et al.*, 2011; Kraemer and Frank, 2010) in order to result in clear guidance to clinicians and patients.

### Multiple mediators

For an 'a priori' hypothesized single mediator of T on O, the process of estimating the mediator effect size is described earlier. In exploring for potential mediators after a RCT, a major advantage to the MK approach (MacKinnon, 2008), using Structural Equation Models, has always been the possibility of including multiple possible mediators for evaluation *simultaneously*. However, every element in a Structural Equation Model embodies assumptions that may or may not be true, in most cases, linearity, absence of interactions, and independence of mediators and often, absence of moderators. If the assumptions *all* hold, clearly the Structural Equation Model approach will yield richer information. If *any one* of these assumptions does not hold, the conclusions may not hold at all. Moreover, goodness-of-fit tests can result in rejection of the model, but not rejecting the model is different from demonstrating that the model is a correct representation of the system.

The MA exploratory approach is different. First, all possible mediators are ordered in time during the intervention. Those possible mediators that are coincident (e.g. change during the first two weeks of treatment), are evaluated to see if any are proxy to others, or overlapping with others (for definitions and criteria, see Kraemer *et al.*, 2008). Those that are proxy are removed from consideration. Redundant predictors are removed, perhaps by combining them to obtain a more reliable measure of their common construct. Thus the remaining mediators at each time point are relatively independent of each other, each representing a different path to the outcome. Then the remaining possible mediators at different times are checked to ascertain whether some are proxy to each other, and again, if such are found, these are removed.

The next question is whether some later M's mediate the effect of some earlier M's on outcome. This may identify a chain of mediators, where M1 mediates the effect of T on M2, M3 mediates the effect of M2 on M4, etc., eventually linking with O. Such a chain of mediators can be very important, in that they signal time points when interventions can be modified to strengthen the chain leading to improved outcome. There may be multiple separate chains of mediators. If moderators of T on O were identified, there may be separate subpopulations, in which the chains of mediators are different. This approach is far more cumbersome than proposing a Structural Equation Model, but is less dependent on assumptions made, and more likely to reveal true complexity (for an example of this approach in an observational study, see Essex *et al.*, 2006).

The Structural Equation Model approach can be likened to proposing a completed picture and then asking whether that picture as a whole is consistent with the data or not. In contrast, the MA approach is like creating a jigsaw puzzle picture by examining data, making sure that all the pieces belong to one picture (absence of moderators) and separating them if they do not, discarding irrelevant or redundant pieces (overlapping or proxy variables), and then fitting the remaining pieces together to form one or more pictures. Since both approaches are hypothesis-generating (exploratory) rather than hypothesis-testing, in either case, any hypotheses so developed would have to be validated by formal hypothesis testing in an independent study designed specifically for that purpose.

## Discussion

The objective here was to develop a mediator effect size in RCTs minimally restricted by any linear model assumptions, not to reject those assumptions but to allow development of methods to address these issues whether or not those assumptions are met.

One goal is to separate the direct and indirect effects of T on O. The results of dissecting the Overall SRD into one


Int. J. Methods Psychiatr. Res. 23(4): 401–410 (2014). DOI: 10.1002/mpr

component that applies when M and T are not correlated (Null SRD) and another then applies when they are correlated (Overall SRD − Null SRD) accomplishes this purpose.

In medical research, to unequivocally show that M is a *causal* mediator of T on O using RCT methodology, one would have to structure a new treatment that incorporates all that T1 offers (T1 here is the overall preferred treatment), and add a component that would increase the beneficial effect of the mediator M on O, either by changing the correlation of T and M, or by changing the impact of M on O, a new treatment say T1+. Then a RCT comparing T1+ with T1 results in a SRD comparing the two would indicate a causal role of M in affecting O. If the SRD comparing T1+ with T1 is zero, then M has no causal effect, whatever the value of MedES relating M, T (T1 versus T2), O. When the linear model criteria hold, conclusions based on both approaches will be concordant. However, as noted by Judd *et al.* (2001, page 133) concerning the linear model approach: "… these analyses make sense only if the many assumptions that underlie them can realistically be made." In contrast, the MA approach is more rigid in the criteria it imposes, but allows flexibility in how to demonstrate that the criteria defining moderators and mediators apply, which then allows researchers to use the linear model where it is appropriate and to develop other models where it is not.

Finally, an unanswered question, and perhaps an unanswerable one, is: How big a MedES in a RCT is big enough for clinical significance? The answer depends strongly on the context, the vulnerability of the population, the costs and risks of T1 and T2, and the impact on patients of inadequate treatment. A relatively small MedES might be more important if the treatment were for cancer and the outcome early death, than if the treatment were for the common cold and the outcome shorter duration of cold symptoms. Moreover, two mediators of the same T on the same O in the same population having the same MedES might have different clinical significance if one could be easily manipulated and were causal, and the other were not.

The driving force in the search for moderators/mediators in clinical research is the belief that recognizing moderator/mediators can help make the most informed clinical and public health decisions given limited resources. Understanding which factors or combination of factors identify those most likely to benefit from T1 rather than T2 (moderators) is essential for treatment matching. Understanding what may be the mechanisms (mediators) by which a given treatment has an impact on outcome is fundamental to refining and optimizing treatments. Interpretable effect sizes are vital to this process.

## References


Acion L., Peterson J.J., Temple S., Arndt S. (2006) Probabilistic index: an intuitive non-parametric approach to measuring the size of treatment effects. *Statistics in Medicine*, **25**(4), 591–602.

Baron R.M., Kenny D.A. (1986) The Moderator–Mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, **51**(6), 1173–1182.

Connolly S.J. (2006) Use and misuse of surrogate outcomes in arrhythmia trials. *Circulation*, **113**(6), 764–766.

Efron B., Gong G. (1983) A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, **37**(1), 36–48.

Efron B., Tibshirani R. (1995) Computer-Intensive Statistical Methods, Technical Report No. 174, Stanford, CA, Division of Biostatistics, Stanford University.

Essex M.J., Kraemer H.C., Armstong J.M., Boyce W.T., Goldsmith H.H., Klein M.H., Woodward H., Kupfer D.J. (2006) Exploring risk factors for the emergence of children's mental health problems. *Archives of General Psychiatry* **63**(11), 1246–1256.

Garber A.M., Tunis S.R. (2009) Does comparative-effectivenss research threaten personalized medicine? *The New England Journal of Medicine* **360**(19), 1925–1927.

Goldin P.R., Ziv M., Jazaieri H., Werner K., Kraemer H.C., Heimberg H., Gross J.J. (2012) Cognitive reappraisal self-efficacy mediates the effects of individual cognitive-behavioral therapy for social anxiety disorder. *Journal of Consulting and Clinical Psychology*, **80**(6), 1034–1040.

Grissom R.J. (1994) Probability of the superior outcome of one treatment over another. *Journal of Applied Psychology* **79**(2), 314–316.

Hsu L.M. (2004) Biases of success rate differences shown in binomial effect size displays. *Psychological Bulletin*, **9**(2), 183–197.

Jain K.K. (2002) Personalized medicine. *Current Opinion in Molecular Therapeutics*, **4**(6), 548–558.

Judd C.M., Kenny D.A., McClelland G.H. (2001) Estimating and testing mediation and moderation in within-subjects designs. *Psychological Methods* **6**(2), 115–134.

King A.C., Ahn D.F., Atienza A.A., Kraemer H.C. (2008) Exploring refinements in targeted behavioral medical intervention to advance public health. *Annals of Behavioral Medicine*, **35**(3), 251–260.

Kraemer H.C. (2008) Toward non-parametric and clinically meaningful moderators and mediators. *Statistics in Medicine*, **27**(10), 1679–1692.

Kraemer H.C., Blasey C. (2004) Centring in regression analysis: a strategy to prevent errors in statistical inference. *International Journal of Methods in Psychiatric Research*, **13**(3), 141–151.

Kraemer H.C., Frank E. (2010) Evaluation of comparative treatment trials: assessing the clinical benefits and risks for patients, rather than statistical effects on measures. *Journal of the American Medical Association*, **304**(6), 1–2.

Kraemer H.C., Frank E., Kupfer D.J. (2006) Moderators of treatment outcomes: clinical, research, and policy importance. *Journal of the American Medical Association*, **296**(10), 1–4.

Kraemer H.C., Frank E., Kupfer D.J. (2011) How to assess the clinical impact of treatments on patients, rather than the statistical impact of treatments on measures. *International Journal of Methods in Psychiatric Research*, **20**(2), 63–71.

Kraemer H.C., Kiernan M., Essex M.J., Kupfer D.J. (2008) How and why criteria defining moderators and mediators differ between the Baron & Kenny and MacArthur approaches. *Health Psychology*, **27**(2), S101–S108.

Kraemer H.C., Kupfer D.J. (2006) Size of treatment effects and their importance to clinical research and practice. *Biological Psychiatry*, **59**(11), 990–996.

Kraemer H.C., Stice E., Kazdin A., Kupfer D. (2001) How do risk factors work together to produce an outcome? Mediators, moderators, independent, overlapping and proxy risk factors. *The American Journal of Psychiatry*, **158**(6), 848–856.

Kraemer H.C., Wilson G.T., Fairburn C.G., Agras W.S. (2002) Mediators and moderators of treatment effects in randomized clinical trials. *Archives of General Psychiatry*, **59**(10), 877–883.

Lesko L.J. (2007) Personalized medicine: elusive dream or imminent reality? *Clinical Pharmacology Therapy*, **81**(6), 807–815.

MacKinnon D.P. (2008) Introduction to Statistical Mediation Analysis, New York: Psychology Press.

MacKinnon D.P., Lockwood C.M., Hoffman J.M., West S.G., Sheets V. (2002) A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, **7**(1), 83–104.

McGraw K.O., Wong S.P. (1992) A common language effect size statistic. *Psychological Bulletin*, **111**(2), 361–365.

Pearl J. (2013) Causality: Models, Reasoning, and Inference, second edition, Cambridge: Cambridge University Press.

Preacher K.J., Kelley K. (2011) Effect size measure for mediation models: quantitative strategies for communicating indirect effects. *Psychological Methods*, **16**(2), 93–115.

Rubin D.B. (2004) Teaching statistical inference for causal effects in experiments and observational studies. *Journal of Educational and Behavioral Statistics*, **29**(3), 343–367.

Silverman W.A. (1998) Where's the Evidence? Debates in Modern Medicine, Oxford: Oxford University Press.

Wilkinson L. (1999) The task force on statistical inference. Statistical methods in psychology journals: guidelines and explanations. *American Psychologist*, **54**(8), 594–604.


Int. J. Methods Psychiatr. Res. 23(4): 401–410 (2014). DOI: 10.1002/mpr