SUPPLEMENTARY INFORMATION

The focus in this paper was on information necessary for clinicians and patients to understand the issues of moderation/mediation in randomized clinical trials and their importance to clinical decision-making. Issues important to clinical researchers in implementing such research were omitted, such as how to estimate SRD or issues related to design and analysis. Some of these issues are here briefly discussed.

SUCCESS RATE DIFFERENCE (SRD)

The Success Rate Difference (SRD) is not the usual treatment effect size used in randomized clinical trials, but it is the most flexible, and easily understood by clinicians.

Binary Outcomes: The proposal for SRD and Number Needed to Treat (NNT) originated with a binary outcome (success/failure), the difference in the success rates in the two treatments. Then the probability that a Treatment patient has outcome better than a Control patient is p1(1-p2), and that a Control patient has outcome better than a Treatment patient is p2(1-p1), where p1 and p2 are respectively the success rates in the two groups. The difference, SRD, is then p1-p2.

The distribution of SRD is that of the difference between two sample binomial proportions, which can be computed exactly or approximated with a normal distribution. The standard error of estimated SRD and a confidence interval for the population SRD can be computed using standard methods. In designing the trial, if the critical value of SRD is c, then the sample size is determined by having at least (say) 80% power to detect the

difference between success rates 0.5-c/2 vs. 0.5+c/2, which would then yield at least 80% power to detect any SRD exceeding c.

The usual effect size here is the Odds Ratio comparing the odds of success in the two groups. However, Odds Ratio is not a clinically interpretable effect size (1-4). For any Odds Ratio (OR) greater than 1, $0<SRD\leq(\sqrt{OR}-1)/(\sqrt{OR}+1)$. Thus the difference in success rates leading to any OR>1 may be arbitrarily close to zero. The Odds Ratio only puts an upper limit on the magnitude of SRD (lower limit on NNT), which does not well guide clinical decision-making. It is for this reason that one cannot design a trial with more than 80% power to detect any OR greater than a set critical value without placing some restrictions.

Normally Distributed Outcomes: If the outcome measures in the two treatment groups are continuous, normally distributed with equal variances, the usual treatment effect size reported is Cohen's d, the mean difference between the two groups divided by the common standard deviation (5). Then $SRD=2\Phi(d/\sqrt{2})-1$ is simply a rescaling of Cohen's d (6). The known distribution of the sample estimate of Cohen's d can be used to estimate the standard error of SRD and a confidence interval for population SRD(7). If the critical effect size of SRD is c, then the power computation is done to have more than (say) 80% power to detect any d greater than $\sqrt{2}\Phi^{-1}((c+1)/2)$, using standard methods(5, 8). The standards for small, medium and large effect sizes, proposed by Cohen are d=.2, .5, .8, and these are the basis for the standards for SRD and NNT mentioned in text.

If the distributions of outcomes are normal but with unequal variances, SRD still equals $2\Phi(d/\sqrt{2})-1$, but with d now defined as the mean difference divided by the square root of the average of the two variances.

Ordinal Outcomes: If the outcome measures in the two treatment groups are ordinal, whatever their distribution, the population SRD can be estimated using the Mann-Whitney U-statistic: SRD=2U/(mn)-1, where m and n are the sample sizes in the two treatment groups. Standard errors and confidence intervals can be based on bootstrap methods, for the distribution of the sample SRD differs depending on the unknown distributions of the outcomes in the two treatment groups. Approximate power computations can be done using the methods currently used for the Mann-Whitney-Wilcoxon test.

Survival as the Outcome: When the outcome is a survival time (possibly censored), the NNT can be estimated using the methods described in (9), and converted to SRD (1/NNT). Power computations are done as usual for the comparison of survival curves.

"Brute Force" Estimates: It should be noted that one can always obtain a sample estimate of the SRD by doing all pairwise comparisons between the outcomes of the m and n patients assigned to the two treatments groups, and standard errors and confidence intervals can be obtained using bootstrap methods. With a trial of any size, this is a tedious and costly procedure. If, for example, there were 100 patients in each treatment group, that would mean 10,000 pairwise comparisons. However, when the outcome is some multivariate descriptor (say a list of benefits and harms for each patient), this may seem the only recourse, and a very difficult one.

One tactic that might simplify the process, however, is that of taking a random sample of, say, 100 paired such outcomes, having these examined by a panel of clinicians, patients and other stake-holders, who use their judgment to declare which of each pair (blinded to treatment choice) is preferable (ties allowed). This allows clinicians or patients to assess how much benefit they view as cancelled by each harm or vice versa(10, 11). A

statistical algorithm might then be developed based on these responses to predict the experts' judgments. If such an algorithm is validated on an independent sample of paired patients, the algorithm can be used to score each patient in the trial (as well as in other trials sampling the same population with the same outcome measures), thus producing an ordinal scale (often approximately normally distributed) in which the simple approaches described above can be used (11).

VALIDATION OF A MODERATOR FOUND IN EXPLORATION

The result of exploration is a hypothesis to be tested in a future independent sample (validation). One important question for clinical researchers is that of how such a validation study might best be designed. What follows is only one possible such answer, with emphasis on clinical importance.

The only moderator of *clinical* importance is one which can be used to divide the population into two subgroups, one in which Treatment is preferred to Control, the other in which Control is preferred to Treatment. If there were a remaining subpopulation in which there is no preference, this subpopulation would be excluded from this validation study. The sample is stratified into the two strata based on the hypothesized moderator(s): Treatment-Preferred and Control-Preferred, and each patient is randomly assigned to the Preferred versus the Non-Preferred treatment determined by his/her value of the moderator.

The study should be designed with adequate power to detect whatever critical value of SRD was used in the original study on which exploration was based. The SRD to be tested and estimated is that comparing the Preferred versus the Non-Preferred treatment,

which should be much greater than that between T1 and T2 (which can also be estimated from the same data).

VALIDATION OF A MEDIATOR FOUND IN EXPLORATION

A similar question arises when a mediator is found in exploration. The *clinical* importance of a mediator lies in the fact that it can be used to improve treatment effects, that is, its value lies in whether the mediator has a causal effect on outcome, and can be exploited to improve treatment outcome, neither of which is assured in exploration. What needs to be done is to structure a new treatment, say Treatment-Augmented. The augmentation of Treatment would be via treatment protocol changes to Treatment that would amplify the contribution of the mediator(s) in the direction of improving outcome.

A randomized clinical trial would then be done comparing Treatment-Augmented vs. the original Treatment and (possibly but not necessarily) the original Control. The crucial effect size now compares Treatment-Augmented with the original Treatment. The critical effect size in designing this study may be smaller than that in comparing the original Treatment versus the original Control, particularly if the original Treatment were more effective than Control.

LINEAR MODEL: The linear models originally proposed by Baron and Kenny for evaluation of moderators and mediators were:

(1) M=b0+b1*T+E

(2) O=c0+c1*T+c2*M+c3*T*M+E*,

where M is the proposed moderator or mediator, T represents here the choice between Treatment and Control. Conditional on the observed T's and M's, E and E* are error terms, assumed to have normal distributions with mean zero. E is assumed to be independent of T with within group variance σ^2_M, and E* to be independent of both T and M with variance $\sigma^2_{e.}$

Because of the assumptions in Equation (1), the proposed mediator M has a normal distribution within each treatment group with equal variances. Thus M can only be a change that occurs during treatment, not an event. Also, since M and E* both have normal distributions within each group, O is also normally distributed within each treatment group

Since, with the interaction in the model, the meaning of the regression coefficients change depending on how T and M are coded(12), for moderator/mediator analysis, it is stipulated that T is coded +1/2 for Treatment and -1/2 for Control. For moderator analysis, M is coded as deviations from the total sample mean pre-randomization. For mediator analysis, M simply measures the change from baseline. Usual Linear Regression approaches are used to test hypotheses. To show that M (a baseline variable in a randomized clinical trial) is a moderator, one needs to show that c3 (the interaction term) is non-zero. To show that M (change during treatment) is a mediator, one needs to show that b1 is non-zero and that either c2 and/or c3 are non-zero.

While null-hypothesis testing using such linear models are often done conditional on the observed values of M, for moderator/mediator analyses, it must be noted that M is also a random variable. With randomization, the population distributions of baseline M are the same in Treatment and Control. Thus if one stratified the population on values of M, all $D_m=0$. Cohen's d within a stratum defined by M=m is $d_m=(c1+c3*m)/\sigma_E$, and thus for each

m, SRD(m,m)= $2\Phi(d_m/\sqrt{2})-1$ which are the same for all values of m if and only if c3 is equal to zero. Hence, when the linearity assumptions hold, the original and the general approaches to moderation arrive at the same answer.

For a moderator (i.e., c3≠0) to have *clinical* importance (i.e., distinguish those patients more likely to respond to Treatment from those more likely to respond to Control), the two regression lines in Equation (2) must cross at M=M* within the range of M observed, i.e., at M*=-2c1/c3. The clinical importance of a moderator is determined by the proportion of the population on either side of M*, and by the relative magnitudes of the effect sizes for patients with M on either side of M*.

For a change that occurs during treatment (M), under the linear model assumptions, treatment choice (T) and M are correlated if and only if b1 is non-zero. The mean value of M in the Treatment group is b0+.5b1, and that in the Control group is b0-.5 b1. The average of those two means is b0 and the half-difference is b1. The variance of M in both groups is σ^2_M.

With Equation (2), the difference between the two outcome means is c1+c2*b1+c3*b0, and the average of the two variances is $\sigma^2_e+(c2^2+.25c3^{2)}\sigma^2_M$. Thus Cohen's d (with unequal variances) for the overall effect of treatment is the ratio of that mean difference divided by the square root of the average variance, Overall-d. The Overall SRD is then $2\ \Phi(\text{Overall-d}/\sqrt{2})-1$. It can then be seen that M mediates the effect of treatment if b1 is non-zero and both c2 and c3 contribute to the effect of treatment (mediate) provided either c2 and/or c3 is non-zero. The effect of mediation on outcome reflects not only an effect on the mean difference but also via an effect on the within-group variances.

Since the only indicator of correlation between T and M under these assumptions is that b1 is not equal to zero, the direct effect (that which would have pertained if the correlation were zero) is here given by the ratio of c1+c3*b0 to the same square root of the average variance, Direct-d, and the Direct-SRD is then 2 Φ(Direct-d/$\sqrt{2}$)-1. The difference between the Overall-SRD and the Direct-SRD is then the mediator effect size under this model. Thus the mediator effect size depends crucially on how strongly M and T are correlated (b1), on the magnitudes of c2 and c3, and on the distribution of M in the population.

However, in practice, a proposed mediator is often not normally distributed in both treatment groups, and even if so, may not have equal variances. The linearity assumptions of Equation 2 may not hold, and even if they do, if M is not normally distributed, the distribution of O is not normal, but a mixture of normal distributions within each treatment group, the mixture determined by the distributions of M. Moreover the original approach to mediators assumed that c3=0, i.e., that the two regression lines for O were parallel. However, assuming that c3=0 when that is not so, can introduce serious error into findings. Moreover, as shown by the effect sizes, an interaction effect, c3, can also contribute to explaining all or part of the association between T and O, either in conjunction with c2 or in absence of c2. Thus if the assumptions of the original model hold, the conclusions of the original approach will be consistent with the MacArthur approach. If not, the MacArthur approach avoids limiting assumptions that may lead to false conclusions.

In a multi-site study with continuous outcome, the most common approach is ANOVA with T (treatment choice), S (site), and the T by S interaction, which is also based on linearity assumptions. Most computer packages perform the centering of S correctly, so

that the T-effect tests the average within group treatment effect size, and the interaction effect tests whether the effect sizes (Cohen's d under the assumptions, and thus SRD) are the same across the sites. If the sample sizes per cell are near equal, the results of such ANOVA tend to be quite robust to deviations from the underlying assumptions. However, if the assumptions of ANOVA are seriously violated, the model in Table 1 can be used directly, and the test of equality of the SRDs across sites can be tested using Bootstrap methods.

There is a more serious problem with a binary outcome. The most common approach is a Logistic Regression analysis with T, S and the T by S interaction. Centering in ANOVA is generally done to assure the interpretation of the effects of interest, but that is not always true with use of the Logistic Regression model (12). Even when centering is handled correctly, the T by S interaction tests whether the Odds Ratios at the different sites are equal. However, the Odds Ratios may be different when the SRDs are not, and the differences in the Odds Ratio may not be detectable even when the SRD differences are large. In this case, the model in Table 1 is better used directly to compare the SRDs from each site.

TOWARD RDoC GOALS

Differential response to treatment (identified by moderators) is a major indicator of individual differences among patients currently labeled as having one diagnosis, particularly if the moderators are genetic, parameters of brain structure/function, or biochemical markers. Identification of such biological moderators should motivate careful assessment of patients recruited into the initial trial to permit assessment of such possible moderators in the exploratory phase to follow a trial.

Similarly the changes in gene expression, parameters of brain structure/function or biochemical markers during treatment, if mediators of treatment outcome, may be very important to understanding the biological processes that underlie a particular diagnosis. Again that should motivate careful assessment of patients during treatment in a trial, to isolate the biological mediators of treatment response.

REFERENCES

1. McGough JJ, Faraone SV. Estimating the size of treatment effects: moving beyond p values. Psychiatry (Edgmont). 2009;6(10):21-9.
2. Newcombe RG. A deficiency of the odds ratio as a measure of effect size. Statistics in Medicine. 2006;25:4235-40.
3. Kraemer HC. Reconsidering the Odds Ratio as a Measure of 2X2 Association in a Population. Statistics in Medicine. 2004;23(2):257-70.
4. Kraemer HC, Kazdin AE, Offord DR, Kessler RC, Jensen PS, Kupfer DJ. Measuring the potency of a risk factor for clinical or policy significance. Psychological Methods. 1999;4(3):257-71.
5. Cohen J. Statistical Power Analysis for the Behavioral Sciences. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
6. Kraemer HC, Kupfer DJ. Size of Treatment Effects and their Importance to Clinical Research and Practice. Biological Psychiatry. 2006;59(11):990-6.
7. Hedges LV, Olkin I. Statistical Methods for Meta-Analysis. Orlando: Academic Press Inc.; 1985.
8. Kraemer HC, Blasey C. How Many Subjects? Statistical Power Analysis in Research (Second Edition). Los Angeles, CA: Sage Publications; 2015.
9. Altman DG, Andersen K. Calculating the number needed to treat for trials where the outcome is time to an event. British Medical Journal. 1999;319:1492-5.
10. Kraemer HC, Frank E, Kupfer DJ. How to assess the clinical impact of treatments on patients, rather than the statistical impact of treatments on measures. Int J Methods Psychiatr Res. 2011; 20(2):63-72.
11. Wallace ML, Frank E, Kraemer HC. A Novel Approach for Developing and Interpreting Treatment Moderator Profiles in Randomized Clinical Trials. JAMA Psychiatry. 2013;70(11):1241-7.
12. Kraemer HC, Blasey C. Centring in Regression Analysis: A Strategy to Prevent Errors in Statistical Inference. International Journal of Methods in Psychiatric Research. 2004;13(3):141-51.