ESTIMATING THE SIZE OF TREATMENT EFFECTS: Moving Beyond *P* Values

by JAMES J. MCGOUGH, MD AND STEPHEN V. FARAONE, PhD

Dr. McGough is Professor of Clinical Psychiatry at the Semel Institute for Neuroscience and Human Behavior and David Geffen School of Medicine at the University of California, Los Angeles. Dr. Faraone is Professor of Psychiatry and Behavioral Sciences at the State University of New York Upstate Medical University, Syracuse.

Psychiatry (Edgemont) 2009;6(10):21–29

ABSTRACT

Objective: To increase understanding of effect size calculations among clinicians who over-rely on interpretations of *P* values in their assessment of the medical literature.

Design: We review five methods of calculating effect sizes: Cohen's *d* (also known as the standardized mean difference)—used in studies that report efficacy in terms of a continuous measurement and calculated from two mean values and their standard deviations; relative risk—the ratio of patients responding to treatment divided by the ratio of patients responding to a different treatment (or placebo), which is particularly useful in prospective clinical trials to assess differences between treatments; odds ratio—used to interpret results of retrospective case-control studies and provide estimates of the risk of side effects by comparing the probability (odds) of an outcome occurring in the presence or absence of a specified condition; number needed to treat—the number of subjects one would expect to treat with agent A to have one more success (or one less failure) than if the same number were treated with agent B; and area under the curve (also known as the drug-placebo response curve)—a six-step process that can be used to assess the effects of medication on both worsening and improvement and the probability that

FUNDING: Funding for the development of this manuscript was provided by Shire Development Inc.

FINANCIAL DISCLOSURE: Dr. McGough has served as a consultant to and received research support from Eli Lilly & Company, Shire Pharmaceuticals, and the National Institutes of Health. In the past year, Dr. Stephen Faraone has received consulting fees and has been on Advisory Boards for Eli Lilly and Shire and has received research support from Eli Lilly, Pfizer, Shire, and the National Institutes of Health. In previous years, Dr. Faraone has received consulting fees or has been on advisory boards or has been a speaker for the following sources: Shire, McNeil, Janssen, Novartis, Pfizer and Eli Lilly. In previous years, he has received research support from Eli Lilly, Shire, Pfizer and the National Institutes of Health.

ADDRESS CORRESPONDENCE TO: James J. McGough, MD, 300 UCLA Medical Plaza, Suite 1414, Los Angeles, CA 90095; E-mail: jmcgough@mednet.ucla.edu; Phone: (310) 794-7841; Fax: (310) 267-0378.

KEY WORDS: Effect sizes, clinical trials, comparing treatment outcomes, pediatric psychopharmacology

a medication-treated subject will have a better outcome than a placebo-treated subject.

Conclusion: Effect size statistics provide a better estimate of treatment effects than *P* values alone.

INTRODUCTION

The purposes of this article are to review methods for comparing the efficacy of interventions across studies and to provide examples of how these methods can be used to make treatment decisions. Significance (the probability that an observed outcome of an experiment or trial is not due to chance alone), direction (positive or negative), magnitude (absolute or relative size), and relevance (the degree to which a result addresses a research topic) are critical elements in the interpretation of treatment effects, but more often than not, clinicians commonly judge results from clinical trials based on their interpretation of *P* values, a measure of statistical significance only. Regardless of the statistic used, when outcomes from two treatments are different at a level of $P<0.05$, it is generally assumed that one treatment, often an active drug, is superior to the other, often a placebo. More specifically, a *P* value indicates the probability, given a particular data set with a particular design and sample size, that differences detected between two treatments are merely due to chance and do not reflect true differences. A *P* value gives a measure of Type I error—the probability of incorrectly detecting a difference between two treatments when no difference exists. A related concept, though less well appreciated, is Type II error—the probability of failing to detect significant differences when in fact they do exist. Careful consideration is required when attempting to balance Type I (erroneous rejection of the null hypothesis) and Type II (erroneous failure to reject the null hypothesis) errors because assumptions based on the topic studied will affect the sample sizes necessary to detect differences between groups.[1] The power of a particular study (defined as 1 minus the specified Type II error) is the probability of finding a true difference between groups if one exists, given a study's sample size and assuming the anticipated magnitude of treatment effect reflects the actual magnitude. Despite attention to these issues as a part of experimental design, an over-reliance on *P* values in the interpretation of clinical trial results can lead either to a naïve acceptance of marginal treatments as efficacious or the rejection of potentially efficacious treatments for a lack of statistical significance.

An issue related to *P* values is the 95-percent confidence interval (CI). A 95-percent CI reveals a range of values around a sample mean in which one can assume, with 95-percent certainty, that the true population mean is found. If the 95-percent CIs of two treatments do not overlap, by definition the means will be significantly different at a level of $P<0.05$.

While a significant *P* value suggests that something nonrandom has occurred, it does not inform us about the clinical significance of the nonrandom effect. For example, the magnitude of statistical significance is heavily influenced by the number of patients studied. Other things being equal, larger samples yield more significant *P* values than smaller samples. Consequently, a large trial of a marginally effective treatment can have greater statistical significance than a small trial of a highly effective treatment. When the results of separate clinical trials are reported as statistically significant, one cannot base decisions about the relative effects of the different treatments by comparing *P* values.[2]

EFFECT SIZE

An effect size is a statistical calculation that can be used to compare the efficacy of different agents by quantifying the size of the difference between treatments. It is a dimensionless measure of the difference in outcomes under two different treatment interventions. Effect sizes thus inform clinicians about the magnitude of treatment effects. Some methods can also indicate whether the difference observed between two treatments is clinically relevant. An effect size estimate provides an interpretable value on the direction and magnitude of an effect of an intervention and allows comparison of results with those of other studies that use comparable measures.[2,3] Interpretation of an effect size, however, still requires evaluation of the meaningfulness of the clinical change and consideration of the study size and the variability of the results. Moreover, similar to statistical significance, effect sizes are also influenced by the study design and random and measurement error. Effect size controls for only one of the many factors that can influence the results of a study, namely differences in variability. The main limitation of effect size estimates is that they can only be used in a meaningful way if there is certainty that compared studies are reasonably similar on study design features that might increase or decrease the effect size. For example, the comparison of effect sizes is questionable if the studies differed substantially on design features that might plausibly influence drug/placebo differences, such as the use of double-blind methodology in one study and non-blinded methodology in the other. It would be impossible to determine whether the difference in effect size was attributable to differences in drug efficacy or differences in methodology. Alternatively, if one of two studies being compared used a highly reliable and well-validated outcome measure while the other used a measure of questionable reliability and validity, these different endpoint outcome measures could also lead to results that would not be meaningful.

Five methods of calculating effect sizes are reviewed: (1) Cohen's *d*, (2) relative risk (RR) ratios, (3)

odds ratios (ORs), (4) number needed to treat (NNT), and (5) area under the curve (AUC). We include examples of each method.

1. COHEN'S *d*

Cohen's *d* is used when studies report efficacy in terms of a continuous measurement, such as a score on a rating scale.[4] Cohen's *d* is also known as the standardized mean difference. It should be noted that the term *effect size* is sometimes used to mean Cohen's *d* in particular and not the several other methods discussed in this review. Of the four critical elements in the interpretation of treatment effects, Cohen's *d* is most useful for assessing magnitude of effects.

Cohen's *d* is calculated from two mean values and their standard deviation (SD). These can be endpoint scores in response to drug A versus drug B, endpoint change scores, or endpoint change scores for drug A versus drug B. Cohen's *d* is computed with the following formula:

Cohen's *d* =

$$\frac{\text{Mean of experimental group} - \text{Mean of control group}}{\text{Pooled SD for entire sample}}$$

A Cohen's *d* score of zero means that the treatment and comparison agent have no differences in effect. A Cohen's *d* greater than zero indicates the degree to which one treatment is more efficacious than the other.[3] A conventional rule is to consider a Cohen's *d* of 0.2 as small, 0.5 as medium, and 0.8 as large.[4] A Cohen's *d* score is frequently accompanied by a confidence interval (CI) so that the reliability of the comparison can be assessed. Calculation of a 95-percent CI around the Cohen's *d* score can facilitate the comparison of effect sizes of different treatments. When effect sizes of similar studies have CIs that do not overlap, this suggests that the Cohen's *d* scores are likely to represent true differences between the studies. Effect sizes that have overlapping CIs suggest that the difference in magnitude of the Cohen's *d* scores may not be statistically significant. While *P* values are used to assess whether or not an effect exists, the use of 95-percent CIs allow for assessment of uncertainty in the magnitude of the effect. Cohen's *d* values are more reliable when used in drug/drug (or placebo) comparisons than in baseline/endpoint comparisons, although they are often used in the latter. It is important to note that the interpretation of Cohen's *d* can be problematic in samples with non-normal distributions or restricted ranges, or if the measurement from which it was derived had unknown reliability.[5]

Another method of interpreting effect sizes, provided by Coe (Table 1), converts Cohen's *d* scores to percentiles.[5] For example, in Table 1, a Cohen's *d* of 0.6 signifies that the mean score of subjects in the experimental group is 0.6 SDs above the mean score of subjects in the control group, and that the mean score of the experimental group exceeds the scores of 73 percent of those in the control group.

To illustrate how Cohen's *d* scores are calculated, data from placebo-controlled studies by The Research Unit on Pediatric Psychopharmacology Anxiety Study Group (RUPP) and by Emslie et al are used to demonstrate the process.[6,7]

In the RUPP study of 128 children and adolescents with anxiety disorders treated with fluvoxamine or placebo, the primary study outcome was the Pediatric Anxiety Rating Scale, a continuous measure.[6] The final mean (±SD) score in the group treated with fluvoxamine was significantly lower than in the placebo group (9.0±7.0 vs. 15.9±5.3; $P<0.001$), a difference of -6.9 (95% CI: -4.6, -9.2), and Cohen's $d=1.1$ (15.9-9.0=6.9; 6.9/6.21 [pooled SD]=1.1). In the Emslie et al study of 96 children and adolescents with depression treated with fluoxetine or placebo, the primary study outcome was the Children's Depression Rating Scale-Revised, another continuous measure.[7] The final mean (±SD) score in the group treated with fluoxetine was significantly lower than in the placebo group (38.4±14.8 vs. 47.1±17.0; $P=0.002$), a difference of -8.7 (95% CI: -15.2, -2.2) and Cohen's $d=0.55$ (47.1-38.4=8.7; 8.7/15.9 [pooled SD]=0.55). Using the conventional approach to interpreting this measure,[4] the RUPP study suggests that fluvoxamine treatment of pediatric anxiety has a large effect size, while the Emslie et al study suggests that fluoxetine treatment of pediatric depression has a medium effect size.[6,7] Of note, the Cohen's *d* score of 0.55 from the Emslie data means that 73 percent of placebo patients had worse scores than the average fluoxetine-treated patient (Table 1). The Cohen's *d* of 1.1 in the RUPP study means that at least 84 percent of the placebo group had worse scores than the average

TABLE 1. Conversion of Cohen's *d* scores to percentiles[5]

COHEN'S *d*	PERCENTAGE OF CONTROL GROUP WHO WOULD HAVE A SCORE BELOW THE AVERAGE SUBJECT IN THE EXPERIMENTAL GROUP
0.0	50
0.1	54
0.2	58
0.3	62
0.4	66
0.5	69
0.6	73
0.7	76
0.8	79
0.9	82
1.0	84
1.2	88
1.4	92
1.6	95
1.8	96
2.0	98
2.5	99
3.0	99.9

TABLE 2. Use of Cohen's *d* in studies of stimulants and nonstimulants in children, adolescents, and adults with ADHD, major depressive episode, and dysthymic disorder*

STUDY	TREATMENT, SUBJECTS, AND SCALES[a]	COHEN'S *d*
Biederman et al, 2003 (2 weeks)[8]	XR-MPH (n=65) vs. placebo (n=71). Children. Conners ADHD/DSM-IV Scale	0.9
Faraone and Schreckengost, 2007 (4 weeks)[9]	Lisdexamfetamine (n=218) vs. placebo (n=72). Children. ADHD Rating Scale	30mg, 1.4 50mg, 1.4 70mg, 1.7
Faraone et al, 2004 (meta-analysis of 6 studies)[10]	IR-MPH (n=139) vs. placebo (n=113). Adults. 4 crossover, 2 parallel studies. ADHD Rating Scale	Crossover, 0.9; Parallel, 0.7
Faraone et al, 2002 (meta-analysis of 4 studies)[11]	IR-MAS (n=108) vs. IR-MPH (n=108). Children and adolescents. IOWA Conners Rating Scale, ADHD Rating Scale, global scale	0.25 (pooled); range, –0.5–1.0
Kelsey et al, 2004 (8 weeks)[12]	Atomoxetine (n=133) vs. placebo (n=64). Children. ADHD Rating Scale	0.7
McGough et al, 2006 (1 week)[13]	Transdermal MPH vs. transdermal placebo (N=79). Children. Crossover study. SKAMP deportment scale	0.9
Michelson et al, 2002 (6 weeks)[14]	Atomoxetine (n=85) vs. placebo (n=85). Children and adolescents. ADHD Rating Scale	0.7
Michelson et al, 2003 (10 weeks)[15]	Study 1, atomoxetine (n=141) vs. placebo (n=139); Study 2, atomoxetine (n=129) vs. placebo (n=127). Adults. Conners Adult ADHD Rating Scale	Study 1, 0.4; Study 2, 0.4
Pelham et al, 2001 (7 days)[16]	OROS-MPH vs. IR-MPH (N=68). Children. Crossover study. IOWA Conners Rating Scale	2.0
Spencer et al, 2002 (12 weeks)[17]	Atomoxetine (n=129) vs. placebo (n=124). Children. ADHD Rating Scale	0.7
Swanson et al, 2003 (7 days)[18]	OROS-MPH or IR-MPH vs. placebo (n=64). Children. Crossover study. IOWA Conners Rating Scale	OROS-MPH, 1.7; IR-MPH, 1.6
Weisler et al, 2006 (4 weeks)[19]	XR-MAS (n=188) vs. placebo (n=60). Adults. ADHD Rating Scale	0.8
Wigal et al, 2005 (3 weeks)[20]	XR-MAS (n=102) vs. atomoxetine (n=101). Children. SKAMP deportment scale	XR-MAS, 1.1; atomoxetine, 0.2
Wilens et al, 2005 (8 weeks)[21]	XR-bupropion (n=81) vs. placebo (n=81). Adults. ADHD Rating Scale	0.6
Wolraich et al, 2001 (1–4 weeks)[22]	OROS-MPH (n=95) or IR-MPH (n=97) vs. placebo (n=90). Children. IOWA Conners Rating Scale	OROS-MPH, 1.1; IR-MPH, 1.0

[a]All were parallel studies unless noted otherwise.
ADHD = attention-deficit/hyperactivity disorder; DSM-IV = Diagnostic and Statistical Manual, Fourth Edition; IOWA = Inattention and Overactivity With Aggression; IR = immediate release; MAS = mixed amphetamine salts; MPH = methylphenidate; OROS = osmotic-release oral system; SKAMP = Swanson, Kotkin, Agler, M-Flynn, and Pelham; XR = extended-release.

fluvoxamine-treated patient.

The growing appreciation of the value of an effect size, such as Cohen's *d* in the interpretation of clinical trials, in contrast to simple reliance on whether or not *P* values are "significant," has led to an increased emphasis on effect sizes in clinical reports. Recognizing the limitations of comparing Cohen's *d* scores across studies with different populations, designs, and outcomes, a summary of studies that report Cohen's *d* in medication trials of attention deficit hyperactivity disorder (ADHD) appears in Table 2.[8–22]

Implications for the clinician. Although Cohen suggested conventions by which to interpret the robustness of clinical effect sizes,[4] values derived for Cohen's *d* do not have direct clinical applicability. For example, although deemed a "small" effect, Cohen's *d* in the range of 0.2 from a study comparing treatment-related mortality rates for two chemotherapies for breast cancer would have far greater clinical consequence than a "large" effect of 0.8 from a study of treatment-related ADHD symptom reduction. More to the point, one might choose a cancer intervention with a better side-effect profile, provided the effect size difference for two treatments, with a concomitant difference in survival, was small; whereas, small treatment effects in behavioral disorders, given the great degree of individual response variability, are not usually meaningful. Estimates of Cohen's *d* mainly serve to maximize successful outcomes in randomized clinical trials by providing a basis for determining study samples sizes needed to provide sufficient power to detect meaningful treatment effects.

2. RELATIVE RISK (RR)

Cohen's *d* is useful for estimating effect sizes from quantitative or dimensional measures. For categorical measures, such as "improved" versus "not improved" or "present" versus "absent," two measures that can be used to assess

effects are RR and OR. Consideration of RR is particularly useful in prospective clinical trials to assess differences in treatments. When attempting to interpret treatment effects, RR can be useful for assessing magnitude, direction, and relevance of effects.

The RR is the ratio of patients improving in a treatment group divided by the probability of patients improving in a different treatment (or placebo) group:

RR= Probability of patients improving on medication / Probability of patients improving on placebo

RR is easy to interpret and consistent with the way in which clinicians generally think. RR ratios can range from zero to infinity. In a study of two treatments, an RR of 1 indicates that outcomes did not differ in the two groups, while an RR of 3 indicates that the treatment group had a threefold greater probability than the placebo group of showing improvement.

In the RUPP anxiety study, 76 percent of the subjects receiving fluvoxamine were treatment responders according to Clinical Global Impressions (CGI) improvement ratings, compared with 29 percent in the placebo group.[6] This yields an RR of 2.6, suggesting that pediatric patients treated with fluvoxamine for anxiety disorders had an almost threefold greater probability of responding than those on placebo. In the pediatric depression study,[7] 56 percent of the fluoxetine-treated subjects versus 33 percent of those on placebo were treatment responders, suggesting an RR of 1.7. These results are consistent with the respective effect sizes derived from each study.

Implications for the clinician. While Cohen's *d* is useful as a basis for the design of clinical trials, calculations of RR provide clinicians with a more intuitively obvious means of assessing treatment efficacy. The statistic is easily calculated from information provided in most clinical reports. It is important to note that RR must be interpreted in the context of the actual probability of the events occurring. For example, if a study showed that the incidence of an adverse event on medication was five percent (0.05) and with placebo it was one percent (0.01), then the RR would be 5.0, meaning that the risk of experiencing the event on the drug would be fivefold greater than on placebo. In assessing response to treatment, a response rate of 90 percent for Treatment A and a response rate of 45 percent for Treatment B yields an RR of 2 (twofold greater probability of responding to A).

RR, which reflects a ratio between two conditions, needs to be differentiated from absolute risk, calculated as a ratio or percentage of an event occurring within one group without any context. Differences in absolute risk are calculated as the ratio or percentage of an event in one group minus the same within a comparison group. In the RUPP example described previously, calculation of absolute risk (76% response in the active group minus 29% in the control group) indicates that 47-percent more subjects responded to fluvoxamine compared with placebo. This conveys information differently than the RR, which, as discussed, suggests a threefold probability of improvement with active treatment.

3. ODDS RATIO (OR)

While RR is an appropriate measure for prospective studies, such as randomized clinical trials or cohort studies, OR is suitable for case-control studies, usually when subjects with a given characteristic are compared with those without the characteristic. An additional benefit of using OR as opposed to RR is that by using the log of OR in statistical modeling, confounding variables can be controlled. Although RR may be easier to understand in terms of evaluating the meaningfulness of differences, OR can be used for the same purpose, albeit in a less intuitive manner. Like RR, OR can be useful for assessing magnitude, direction, and relevance of effects.

An OR is computed as the ratio of two odds: the odds that an event will occur compared with the probability that it will not occur. Specifically, it is as follows:

OR= Proportion with a given trait improved (or worsened)/not improved (or not worsened) / Proportion without a given trait improved (or worsened)/not improved (or not worsened)

For a clinical trial, an OR indicates the increase in the odds of improvement (or worsening) that can be associated with a secondary trait under consideration. An OR can also have an associated CI so that the reliability of the comparison can be assessed.

The OR has particular utility in the interpretation of retrospective case-controlled comparisons. It is less frequently used in the interpretation of randomized, controlled trials, but can be used to assess both positive outcomes (such as improvement) and adverse events. Compared with RR, the OR has value in predicting the likelihood of outcomes with low frequency, such as side effects, or to assess differential treatment effects in subjects with various secondary characteristics, such as sex, age group, or comorbid condition.

A recent study used OR to assess some characteristics of patients with and without a diagnosis of dissociative disorders.[23] The respondents were 231 consecutive admissions to an inner-city outpatient psychiatric clinic. Of the 82 respondents who completed the Dissociative Disorders Interview Schedule, 24 were diagnosed with a dissociative disorder and 58 were not. Of those with dissociative disorders, childhood physical abuse was reported in 17 (71%) and was denied in seven (29%). Of those without dissociative disorders, childhood physical abuse was reported in 17 (29%) and was denied

TABLE 3. Number needed to treat (NNT) analyses of data from 10 studies of children, adolescents, and adults With ADHD

STUDY	TREATMENT, SUBJECTS, AND SCALES[a]	NNT
Biederman et al, 2007 (4 weeks)[27]	LDX (n=214) vs. placebo (n=72). Children. % CGI responders[a]	2
Greenhill et al, 2002 (3 weeks)[28]	MR-MPH (n=155) vs. placebo (n=159). Children and adolescents. % CGI responders[a]	3
Kratochvil et al, 2005 (5 weeks)[29]	Atomoxetine + fluoxetine (n=127) vs. atomoxetine+placebo (n=46). Children and adolescents. % responders: ADHD-RS or CDRS-R scores 1 SD of age and sex norms	ADHD-RS, 12; CDRS-R, 6
Kuperman et al, 2001 (7 weeks)[30]	SR-bupropion (n=11) vs. MPH (n=8) vs. placebo (n=8). Adults. % CGI responders[a]	Bupropion vs. placebo, 3; MPH vs. placebo, 4
Pliszka et al, 2000 (3 weeks)[31]	MAS (n=20) vs. MPH (n=20) vs. placebo (n=18). Children. % CGI responders[a]	MAS vs. placebo, 2; MPH vs. placebo, 3
Reimherr et al, 2007 (4 weeks)[32]	OROS-MPH (n=47) vs. placebo. Crossover study. Adults. % responders: ≥50% improvement in WRAADDS scores	3
Riggs et al, 2004 (12 weeks)[33]	Pemoline (n=34) vs. placebo (n=33). Adolescents. % CGI responders[a]	5
Safren et al, 2005 (15 weeks)[34]	Cognitive-behavioral therapy + ADHD medications (n=16) vs. ADHD medications (n=15). Adults. % CGI responders[a]	2
Spencer et al, 2005 (6 weeks)[35]	MPH (n=104) vs. placebo (n=42). Adults. % responders: % CGI responders[a] + >30% reduction in AISRS scores	2
Weiss et al, 2006 (20 weeks)[35]	DEX (n=23) vs. paroxetine (n=24) vs. placebo (n=26). Adults. % CGI responders[a]	DEX vs. paroxetine, 2; vs. placebo, 2

[a]Patients with CGI ratings of much or very much improved (score 1 or 2).
ADHD=attention-deficit/hyperactivity disorder; ADHD-RS=ADHD Rating Scale IV; AISRS=Adult ADHD Investigator System Report Scale; CDRS-R=Children's Depression Rating Scale–Revised; CGI=Clinical Global Impressions scale; DEX=dextroamphetamine; LDX=lisdexamfetamine; MAS=mixed amphetamine salts; MPH=methylphenidate; MR=modified release; OROS=osmotic-release oral system; SD=standard deviation; SR=sustained release; WRAADDS=Wender-Reimherr Adult Attention Deficit Disorder Scale; XR=extended release.

in 41 (71%; OR: 5.86 [17/7]/[17/41], $P<0.001$).[23] This means that the odds of having suffered childhood physical abuse were sixfold greater for patients with dissociative disorders as compared to those without dissociative disorders.

The OR is frequently used to provide estimates of side effects risk. In the Preschool ADHD Treatment Study, children were treated with methylphenidate and assessed for side effects that commonly occur with stimulants.[24] Although preliminary and requiring replication, results suggested that patient genotypes predicted risk for developing certain side effects. For example, the odds of manifesting picking behaviors were threefold greater for children who were homozygous for the dopamine receptor D4 variable-number tandem repeat (DRD4-VNTR) polymorphism four-repeat allele (OR: 3.0; 95% CI: 1.14, 8.14; $P=0.03$) as compared to children with other genotypes, while the odds of developing social withdrawal with increasing dose were almost fourfold greater for children with any copy of the seven-repeat allele (OR: 3.92; 95% CI: 1.20, 12.86; $P=0.01$).[24]

Implications for the clinician. ORs are useful for making inferences about the risk or probability of one outcome versus another, whether these are rates of positive response or some adverse outcome. While an OR might be statistically significant, the increased risk for differences in outcome between groups might be quite small and lack clinical significance if multiple variables moderate response. ORs are most useful in assessing characteristics associated with low-frequency outcomes, such as differential responses between various patient groups or in assessing side effects. ORs do not easily provide estimates of clinical effect size comparable to Cohen's *d* and do not provide a reasonable basis for sample size estimates in clinical trials. When available, ORs provide additional information for predicting risk versus benefit in individual patients.

4. NUMBER NEEDED TO TREAT (NNT)

The NNT is defined as the number of subjects one would expect to treat with agent A to have one more success (or one less failure) than if the same number were treated with agent B. The NNT has been described as the effect size estimate "that seems to best reflect clinical significance...for binary (success/failure) outcomes."[2] NNT is a measure related to absolute risk reduction and may be most useful in assessing relevance of treatment effects. The NNT is computed as follows:

NNT = 100 divided by % improved on treatment - % improved on placebo

In the RUPP anxiety study, 76 percent of the subjects receiving fluvoxamine versus 29 percent of the

placebo group were treatment responders.[6] This yields NNT=2.1 (100/[76-29]), suggesting that for every two patients treated with active medication, at least one will have a better outcome than if treated with placebo. In the fluoxetine depression study, 56 percent of the subjects receiving fluoxetine versus 33 percent of the placebo group were treatment responders,[7] yielding NNT=4.3(100/[56-33]). This suggests that it is necessary to treat four patients with active medication to get one better response than treatment with placebo. These two NNT values are consistent with previously described effect sizes and RRs for improvement.

NNT is also useful for assessing the risk of negative outcomes. In another study of adolescents with depression, suicide-related events were reported in 4.5 percent of subjects treated with fluoxetine and in 2.7 percent of those taking placebo.[25] This yields NNT=56(100/[4.5-2.7]), suggesting that 56 patients would need to be treated with active medication to observe one additional patient evidencing suicidality compared with placebo.

NNTs for 10 studies of ADHD treatment in children, adolescents, and adults are shown in Table 3.[26–35]

Implications for the clinician. NNT is an easily calculated statistic that is readily interpretable and can provide guidance as to the likelihood of positive or negative outcomes in actively treated versus placebo groups.

5. AREA UNDER THE CURVE (AUC)

The method generally known as the AUC or "area under the ROC curve" (ROC standing for "receiver operating characteristic")[2] has also been described as the "drug-placebo response curve," a generalization of the ROC curve.[36] This measure is most useful for assessing relevance of treatment effects.

There are six steps to developing a drug-placebo response curve:[36] (1) choose an outcome variable, such as the change in a symptom score from baseline to endpoint; (2) for the drug and placebo groups separately, calculate for each observed score the proportion of patients having that score or a better score; (3) for each observed score, plot the proportions computed in step 2 for the drug group on the vertical axis and for the placebo group on the horizontal axis; (4) connect the plotted points and label those that correspond to the best response as the 25th percentile, the median response as the 75th percentile, and the worst response; (5) if the outcome variable is a change score, also label the point corresponding to no change; and (6) plot the diagonal line of no effect (no difference between the drug and placebo groups).

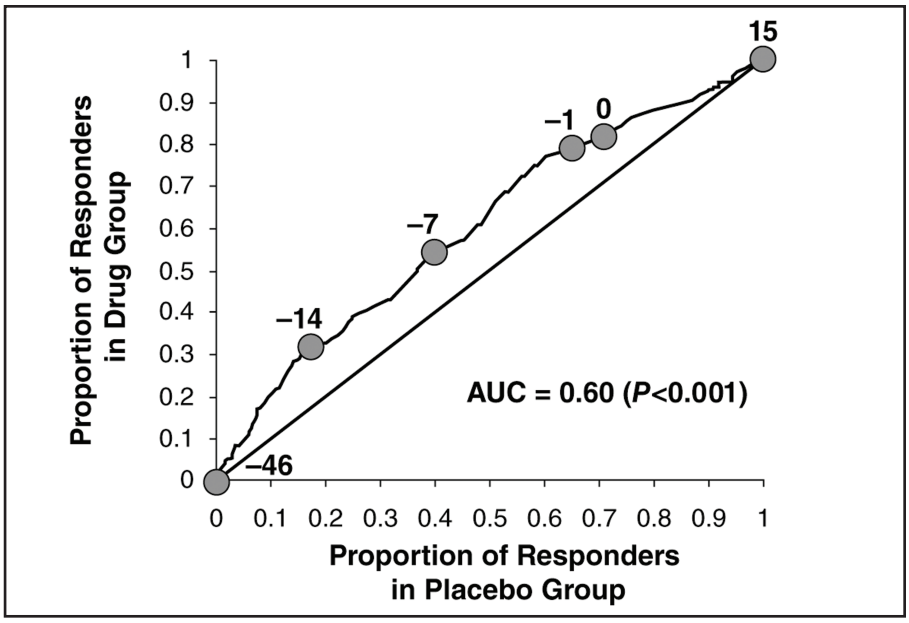

FIGURE 1. Drug-placebo response curve of total Conners Adult ADHD Rating Scale scores in patients receiving atomoxetine or placebo. Used with permission from Biomed Central. Faraone SV, et al. *Behav Brain Funct.* 2005;1:16.[38]

This approach has been applied to data from two studies of atomoxetine versus placebo in adults with ADHD.[37] Subjects (N=267 in study 1 and N=248 in study 2) received 60 to 120mg of atomoxetine or placebo daily for 10 weeks and were assessed with a variety of categorical and quantitative outcomes. In both studies atomoxetine was significantly more efficacious than placebo according to investigator and patient responses on the Conners Adult ADHD Rating Scale (CAARS) (P=0.002 and 0.008, respectively).[37]

Figure 1, based on previous analyses of investigator-rated total CAARS scores,[37] illustrates the application of the drug-placebo response curve to identify additional differences in treatment groups. In this figure, atomoxetine is shown to have produced a drug-placebo response curve that is always above the diagonal, indicating that atomoxetine was more efficacious than placebo throughout the full range of outcome scores. The AUC of 0.60 indicates that 60 percent of subjects treated with atomoxetine showed symptom improvement, while only approximately 40 percent of those treated with placebo improved. The point in the figure labeled –7 is located at coordinates 0.54 and 0.4, indicating that a reduction of –7 in CAARS total scores was seen in 54 percent of patients treated with atomoxetine and in 40 percent of those given placebo. If clinical improvement is defined as a change score on the CAARS between –1 and –14, the figure shows that a majority of patients judged as responsive received atomoxetine rather than placebo. A value of zero indicates no change in CAARS scores. Among the patients receiving atomoxetine, 82

percent had a score of ≥0, indicating that 18 percent experienced a worsening of symptoms. In contrast, 71 percent of placebo patients had a score of ≥0 and thus 29 percent experienced a worsening of symptoms.[37]

Implications for the clinician. Although ROC methodology is fairly complex, it is instructive in two regards. First, as shown in the above example, the method highlights the effects of medication on both worsening and improvement. Although reports from clinical trials focus on the latter, it is of equal importance to know if the placebo group shows more worsening than the treatment group. This helps clinicians when they weigh the pros and cons of treatment. Although the choice to not treat with medication is sometimes justified on the basis of avoiding adverse events, not treating has the adverse effect of worsening outcome in many cases. Another value of ROC is that the AUC statistic can be interpreted as the probability that a patient treated with medication will have a better outcome than a placebo-treated patient. This can be useful when communicating outcome probabilities to patients or their parents.

CONCLUSION

Several statistical methods have been described to help evaluate the magnitude of the differences between two or more interventions in clinical studies and that provide a better estimate of treatment effects than simple reliance on *P* values. These include Cohen's *d* (also known as the standard mean difference), RR, OR, NNT, and AUC (also known as the drug-placebo response curve). None of these methods alone can be used to assess all four critical elements in the interpretation of treatment effects (significance, direction, magnitude, and relevance) and, therefore, a perfect measure of effect size does not exist. It is incumbent on the clinician to put these pieces of information together appropriately. Toward that end, more general use of these methods in clinical research could result in better designed and more relevant studies. Clinicians need to bear in mind that effect size estimates from clinical trials, usually based on active drug versus placebo comparisons, do not correlate directly with patient responses in real-world clinical practice. Nonetheless, a greater awareness of these approaches among clinicians would result in a more nuanced appreciation of the clinical trials literature than reliance on *P* values alone. In turn, this knowledge of treatment effect estimation may be used to promote patient compliance through the use of epidemiological evidence describing the risks of behaviors or the benefits of treatments or interventions.

ACKNOWLEDGMENTS

Editorial assistance was provided by Timothy Coffey, Robert Gregory, Rosa Real, and William Perlman, Excerpta Medica, Bridgewater, New Jersey.

REFERENCES

1. Schulz KF, Grimes DA. The Lancet Handbook of Essential Concepts in Clinical Research. New York (NY): Elsevier Limited;2006.
2. Kraemer HC, Kupfer DJ. Size of treatment effects and their importance to clinical research and practice. *Biol Psychiatry*. 2006;59:990–996.
3. Faraone SV. Understanding the effect size of ADHD medications: implications for clinical care. *Medscape Psychiatry & Mental Health*. 2003;8(2).
4. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*, Second Edition. Hillsdale (NJ): Eribaum;1988.
5. Coe R. It's the effect size, stupid: what effect size is and why it is important. Presented at British Educational Research Association Annual Conference; 2002 Sept 12–14; Exeter, England.
6. The Research Unit on Pediatric Psychopharmacology Anxiety Study Group. Fluvoxamine for the treatment of anxiety disorders in children and adolescents. *N Engl J Med*. 2001;344:1279–1285.
7. Emslie GJ, Rush AJ, Weinberg WA, et al. A double-blind, randomized, placebo-controlled trial of fluoxetine in children and adolescents with depression. *Arch Gen Psychiatry*. 1997;54:1031–1037.
8. Biederman J, Quinn D, Weiss M, et al. Efficacy and safety of Ritalin® LA™, a new, once daily, extended-release dosage form of methylphenidate, in children with attention deficit hyperactivity disorder. *Pediatr Drugs*. 2003;5:833–841.
9. Faraone SV, Schreckengost J. Lisdexamfetamine dimesylate effect size in children with attention-deficit/hyperactivity disorder. Presented at 54th Annual Meeting of the American Academy of Child & Adolescent Psychiatry; 2007 Oct 23-–28; Boston.
10. Faraone SV, Spencer T, Aleardi M, et al. Meta-analysis of the efficacy of methylphenidate for treating adult attention-deficit/hyperactivity disorder. J *Clin Psychopharmacol*. 2004;24:24–29.
11. Faraone SV, Biederman J, Roe C. Comparative efficacy of Adderall and methylphenidate in attention-deficit/hyperactivity disorder: a meta-analysis. *J Clin Psychopharmacol*. 2002;22:468–473.
12. Kelsey DK, Sumner CR, Casat CD, et al. Once-daily atomoxetine treatment of children with attention-deficit/hyperactivity disorder, including an assessment of evening and morning behavior: a double-blind, placebo-controlled trial. *Pediatrics*. 2004;114:e1–e8.
13. McGough JJ, Wigal SB, Abikoff H, et al. A randomized, double-blind, placebo-controlled, laboratory classroom assessment of methylphenidate transdermal system in children with ADHD. *J Atten Disord*. 2006;9:476–485.
14. Michelson D, Allen AJ, Busner J, et al. Once-daily atomoxetine treatment for children and adolescents with attention deficit hyperactivity disorder: a

randomized, placebo-controlled study. *Am J Psychiatry*. 2002;159:1896–1901.
15. Michelson D, Adler L, Spencer T, et al. Atomoxetine in adults with ADHD: two randomized, placebo-controlled studies. *Biol Psychiatry*. 2003;53:112–120.
16. Pelham WE, Gnagy EM, Burrows-Maclean L, et al. Once-a-day Concerta methylphenidate versus three-times-daily methylphenidate in laboratory and natural settings. *Pediatrics*. 2001;107(6).
17. Spencer T, Heiligenstein JH, Biederman J, et al. Results from 2 proof-of-concept, placebo-controlled studies of atomoxetine in children with attention-deficit/hyperactivity disorder. *J Clin Psychiatry*. 2002;63:1140–1147.
18. Swanson J, Gupta S, Lam A, et al. Development of a new once-a-day formulation of methylphenidate for the treatment of attention-deficit/hyperactivity disorder. *Arch Gen Psychiatry*. 2003;60:204–211.
19. Weisler RH, Biederman J, Spencer TJ, et al; SLI381.303 Study Group. Mixed amphetamine salts extended-release in the treatment of adult ADHD: a randomized, controlled trial. *CNS Spectr*. 2006;11:625–639.
20. Wigal SB, McGough JJ, McCracken JT, et al; SLI381.404 Study Group. A laboratory school comparison of mixed amphetamine salts extended release (Adderall XR®) and atomoxetine (Strattera®) in school-aged children with attention deficit/hyperactivity disorder. *J Atten Disord*. 2005;9:275–289.
21. Wilens TE, Haight BR, Horrigan JP, et al. Bupropion XL in adults with attention-deficit/hyperactivity disorder: a randomized, placebo-controlled study. *Biol Psychiatry*. 2005;57:793–801.
22. Wolraich ML, Greenhill LL, Pelham W, et al; Concerta Study Group. Randomized, controlled trial of OROS methylphenidate once a day in children with attention-deficit/hyperactivity disorder. *Pediatrics*. 2001;108:883–892.
23. Foote B, Smolin Y, Kaplan M, et al. Prevalence of dissociative disorders in psychiatric outpatients. *Am J Psychiatry*. 2006;163:623–629.
24. McGough J, McCracken J, Swanson J, et al. Pharmacogenetics of methylphenidate response in preschoolers with ADHD. *J Am Acad Child Adolesc Psychiatry*. 2006;45:1314–1322.
25. Emslie G, Kratochvil C, Vitiello B, et al; Columbia Suicidality Classification Group; TADS Team. Treatment for Adolescents with Depression Study (TADS): safety results. *J Am Acad Child Adolesc Psychiatry*. 2006;45:1440–1455.
26. Biederman J, Krishnan S, Zhang Y, et al. Efficacy and tolerability of lisdexamfetamine dimesylate (NRP-104) in children with attention-deficit/hyperactivity disorder: a phase III, multicenter, randomized, double-blind, forced-dose, parallel-group study. *Clin Ther*. 2007;29:450–463.
27. Greenhill LL, Findling RL, Swanson JM; MPH MR ADHD Study Group. A double-blind, placebo-controlled study of modified-release methylphenidate in children with attention-deficit/hyperactivity disorder. *Pediatrics*. 2002;109(3).
28. Kratochvil CJ, Newcorn JH, Arnold E, et al. Atomoxetine alone or combined with fluoxetine for treating ADHD with comorbid depressive or anxiety symptoms. *J Am Acad Child Adolesc Psychiatry*. 2005;44:915–924.
29. Kuperman S, Perry PJ, Gaffney GR, et al. Bupropion SR vs. methylphenidate vs. placebo for attention deficit hyperactivity disorder in adults. *Ann Clin Psychiatry*. 2001;13:129–134.
30. Pliszka SR, Browne RG, Olvera RL, Wynne SK. A double-blind, placebo-controlled study of Adderall and methylphenidate in the treatment of attention-deficit/hyperactivity disorder. *J Am Acad Child Adolesc Psychiatry*. 2000;39:619–626.
31. Reimherr FW, Williams ED, Strong RE, et al. A double-blind, placebo-controlled, crossover study of osmotic release oral system methylphenidate in adults with ADHD with assessment of oppositional and emotional dimensions of the disorder. *J Clin Psychiatry*. 2007;68:93–101.
32. Riggs PD, Hall SK, Mikulich-Gilbertson SK, et al. A randomized controlled trial of pemoline for attention-deficit/hyperactivity disorder in substance-abusing adolescents. *J Am Acad Child Adolesc Psychiatry*. 2004;43:420–429.
33. Safren SA, Otto MW, Sprich S, et al. Cognitive-behavioral therapy for ADHD in medication-treated adults with continued symptoms. *Behav Res Ther*. 2005;43:831–842.
34. Spencer T, Biederman J, Wilens T, et al. A large, double-blind, randomized clinical trial of methylphenidate in the treatment of adults with attention-deficit/hyperactivity disorder. *Biol Psychiatry*. 2005;57:456–463.
35. Weiss M, Hechtman L; Adult ADHD Research Group. A randomized double-blind trial of paroxetine and/or dextroamphetamine and problem-focused therapy for attention-deficit/hyperactivity disorder in adults. *J Clin Psychiatry*. 2006;67:611–619.
36. Faraone SV, Biederman J, Spencer TJ, Wilens TE. The drug-placebo response curve: a new method for assessing drug effects in clinical trials. *J Clin Psychopharmacol*. 2000;20:673–679.
37. Faraone SV, Biederman J, Spencer T, et al. Efficacy of atomoxetine in adult attention-deficit/hyperactivity disorder: a drug-placebo response curve analysis. *Behav Brain Funct*. 2005;1:16.