Clinical Relevance of the Primary Findings of the MTA: Success Rates Based on Severity of ADHD and ODD Symptoms at the End of Treatment

JAMES M. SWANSON, PH.D., HELENA C. KRAEMER, PH.D., STEPHEN P. HINSHAW, PH.D., L. EUGENE ARNOLD, M.D., C. KEITH CONNERS, PH.D., HOWARD B. ABIKOFF, PH.D., WALTER CLEVENGER, B.A., MARK DAVIES, M.P.H., GLEN R. ELLIOTT, PH.D., M.D., LAURENCE L. GREENHILL, M.D., LILY HECHTMAN, M.D., BETSY HOZA, PH.D., PETER S. JENSEN, M.D., JOHN S. MARCH, M.D., JEFFREY H. NEWCORN, M.D., ELIZABETH B. OWENS, PH.D., WILLIAM E. PELHAM, PH.D., ELLEN SCHILLER, PH.D., JOANNE B. SEVERE, M.S., STEVE SIMPSON, M.A., BENEDETTO VITIELLO, M.D., KAREN WELLS, PH.D., TIMOTHY WIGAL, PH.D., AND MIN WU, M.S.

ABSTRACT

Objectives: To develop a categorical outcome measure related to clinical decisions and to perform secondary analyses to supplement the primary analyses of the NIMH Collaborative Multisite Multimodal Treatment Study of Children With Attention-Deficit/Hyperactivity Disorder (MTA). **Method:** End-of-treatment status was summarized by averaging the parent and teacher ratings of attention-deficit/hyperactivity disorder and oppositional defiant disorder symptoms on the Swanson, Nolan, and Pelham, version IV (SNAP-IV) scale, and low symptom-severity ("Just a Little") on this continuous measure was set as a clinical cutoff to form a categorical outcome measure reflecting successful treatment. Three orthogonal comparisons of the treatment groups (combined treatment [Comb], medication management [MedMgt], behavioral treatment [Beh], and community comparison [CC]) evaluated hypotheses about the MTA medication algorithm ("Comb + MedMgt versus Beh + CC"), multimodality superiority ("Comb versus MedMgt"), and psychosocial substitution ("Beh versus CC"). **Results:** The summary of SNAP-IV ratings across sources and domains increased the precision of measurement by 30%. The secondary analyses of group differences in success rates (Comb = 68%; MedMgt = 56%; Beh = 34%; CC = 25%) confirmed the large effect of the MTA medication algorithm and a smaller effect of multimodality superiority, which was now statistically significant ($p < .05$). The psychosocial substitution effect remained negligible and nonsignificant. **Conclusion:** These secondary analyses confirm the primary findings and clarify clinical decisions about the choice between multimodal and unimodal treatment with medication. *J. Am. Acad. Child Adolesc. Psychiatry*, 2001, 40(2):168–179. **Key Words:** SNAP-IV ratings, orthogonal comparisons, logistic regression, multimodal superiority, clinical relevance.

As described in detail elsewhere (MTA Cooperative Group, 1999a,b), the NIMH Collaborative Multisite Multimodal Treatment Study of Children With Attention-Deficit/Hyperactivity Disorder (MTA) was a randomized clinical trial of four treatment strategies: medication management (MedMgt), behavioral treatment (Beh), the combination of these two (Comb), and an active control condition based on usual treatment available in the community (CC). The investigators (MTA Cooperative Group, 1999a) set statistical power based on ". . . comparisons of core ADHD symptoms between any two treatment arms (critical effect size, 0.4; power, 0.81, with a 5% two-tailed test)." This determined the targeted sample size ($n = 576$) of the study. In the primary analyses, an intent-to-treat (ITT) design was used to evaluate status at 0, 3, 9, and 14 months after treatments were initiated. A random-effects regression (RR) procedure was used to evaluate the effects of treatment,

Accepted August 15, 2000.

Dr. Swanson is Professor of Pediatrics, University of California at Irvine, and Senior Fellow of the Sackler Institute, Cornell Medical Center, New York. Mr. Davies is Research Scientist (Statistician), New York State Psychiatric Institute; Ms. Wu is a Ph.D. candidate in the Division of Biostatistics, School of Public Health, Columbia University; Mr. Clevenger is a computer specialist at UC Irvine; Mr. Simpson is a school psychologist at UC Irvine; and Dr. Owens is a psychologist at UC Berkeley. Other authors' affiliations are listed in the MTA Cooperative Group acknowledgment that appears at the end of the text.

Correspondence to Dr. Swanson, University of California at Irvine, Child Development Center, 19722 MacArthur Blvd., Irvine, CA 92612; e-mail: jmswanso@uci.edu.

time, and site and the interactions of these factors. Anticipating differential impact of treatments across 6 domains and 3 sources, 19 outcome variables (selected by principal component analysis to eliminate redundant measures in the assessment battery) were analyzed.

For 10 of the 19 marker variables, there were statistically significant differential changes over time due to treatment. Paired comparisons of group means (i.e., the average slope of outcome-time regression equations estimated for each individual) were used to interpret the overall pattern of group differences statistically significant at $p < .05$ (after adjustment for multiple comparisons). On core symptoms of attention-deficit/hyperactivity disorder (ADHD), these comparisons suggested the following answers to the three principal questions of the study (MTA Cooperative Group, 1999a): (1) Unimodal comparison: "robust differences" showed that MedMgt produced significantly greater improvement than Beh. (2) Multimodal comparisons: Comb produced significantly greater improvement than Beh, but Comb and MedMgt "did not differ significantly across any domain." (3) Community comparisons: Comb and MedMgt were "generally superior" to CC but Beh was not. This pattern of statistically significant treatment effects can be summarized succinctly by the following relationships among groups: Comb ~ MedMgt > Beh ~ CC. The ">" symbol denotes the point at which some paired comparisons of groups were statistically significant, and the "~" symbol denotes sets of paired comparisons that were not statistically different in the primary analyses. For the symptoms of oppositional defiant disorder (ODD), the same general pattern of statistical significance held, but only for $p < .05$ before adjustments for multiple comparisons.

Well-established statistical procedures were used for the initial analyses of the MTA, and the findings reported by the MTA Cooperative Group (1999a) stand as the primary results of our randomized clinical trial. Taylor (1999) offered praise for the methods and execution of the MTA, which was described as ". . . a landmark in the evolution of children's mental health into an evidence-based discipline." However, the interpretations of the primary findings are complex and may be confusing. For example, it may be difficult to interpret the clinical relevance of statistically significant differences unless the group differences in slopes from outcome-time regression equations are translated into status at the end of treatment. Also, it may be difficult to interpret the nonsignificant differences between treatment groups. For example, even some MTA investigators (National Institutes of Health, 2000; Pelham, 1999) have expressed concerns that a highly publicized finding of the primary analyses—the absence of a statistically significant difference between the Comb and MedMgt groups—may be misinterpreted and lead others to conclude that the use of medication alone is just as effective as multimodal treatment. Similarly, the absence of a statistically significant difference between the Beh and CC groups may be mistakenly cited as evidence of noneffectiveness of behavioral interventions.

Such conclusions would be based on inappropriate acceptance of the null hypothesis on the basis of statistical nonsignificance, which may discount *clinically* significant effects. This was noted in the initial report of the primary findings (MTA Cooperative Group, 1999a): ". . . the chance is high of declaring effect sizes lower than 0.4 not statistically significant, even though some clinicians might regard such differences as clinically significant." In this report we provide information on effect size to emphasize clinical significance rather than statistical significance.

We take four methodological approaches to supplement the primary analyses: (1) We use a narrow measure that summarizes the severity of ADHD and ODD symptoms at the end of treatment instead of a broad set of outcome measures that assess change over time. (2) We perform orthogonal comparisons to compare outcomes of the treatment groups instead of paired comparisons that do not separate significant findings into independent, nonredundant components. (3) We emphasize a categorical outcome measure to measure success rates of the MTA treatments instead of a continuous outcome measure that reflects average response of the groups. (4) We highlight effect size instead of statistical significance to emphasize clinical relevance of the MTA findings.

METHOD

Subjects and Procedures

Methods described elsewhere (Hinshaw et al., 1997) were used to recruit 579 children with ADHD-combined type from geographic locations spread across North America (New York City [Columbia University]; Irvine, CA [UC Irvine]; Pittsburgh [Western Psychiatric Institute and Clinic]; Berkeley, CA [UC Berkeley]; Durham, NC [Duke University]; Queens, NY [Long Island Jewish Medical Center]/Montreal, Quebec [Montreal Children's Hospital]). Each of these subjects was randomly assigned to one of the four groups (Comb, MedMgt, Beh, or CC), and those in the first three groups received treatment for 14 months, delivered by a staff specifically trained for this study and monitored by the MTA investigators. Assessments were performed before, during (at 3 and 9 months), and at the end of treatment.

Key Methodological Issues

Narrow and Broad Outcome Measures. The broad MTA assessment battery (Hinshaw et al., 1997) was designed to cover multiple domains and sources of information, but it included so many instruments that the clinical use of the full battery is impractical (Swanson et al., 1999). Two statistical procedures of the primary analyses were used to reduce this large set (MTA Cooperative Group, 1999a). First, principal component analysis of more than 100 baseline measures identified a small subset (19 key marker variables) that best represents the source by domain assessment strategy. Second, random regression analyses of outcome identified statistically significant effects of treatment × time for 10 of these 19 marker variables. One instrument completed by two sources—the Swanson, Nolan, and Pelham, version IV (SNAP-IV) parent and teacher rating scale (Gaub and Carlson, 1997; Swanson, 1992; Swanson et al., 1983, 1999)—provided 6 of the 10 significant marker variables. The items of the SNAP-IV (Fig. 1) include the 18 ADHD and 8 ODD symptoms specified in the *DSM-IV* (American Psychiatric Association, 1994) and *ICD-10 Classification of Mental and Behavioral Disorders* (World Health Organization, 1992, 1993) (Swanson et al., 1998). These symptoms are scored by assigning a severity estimate for each symptom on a 4-point scale (i.e., 0 = not at all, 1 = just a little, 2 = pretty much, and 3 = very much). For both par-

The MTA SNAP-IV Teacher and Parent Rating Scale
James M. Swanson, Ph.D., University of California, Irvine, CA 92715

Name:_____ Gender:_____ Age:_____ Grade:_____

Ethnicity (circle one which best applies): African-American Asian Caucasian Hispanic Other_____

Completed by:_____ Type of Class:_____ Class size:_____

For each item, check the column which best describes this child:	Not At All	Just A Little	Pretty Much	Very Much
1. Fails to give close attention to details or makes careless mistakes in schoolwork or tasks				
2. Has difficulty sustaining attention in tasks or play activities				
3. Does not seem to listen when spoken to directly				
4. Does not follow through on instructions and fails to finish schoolwork, chores, or duties				
5. Has difficulty organizing tasks and activities				
6. Avoids, dislikes, or reluctantly engages in tasks requiring sustained mental effort				
7. Loses things necessary for activities (e.g., toys, school assignments, pencils, or books)				
8. Is distracted by extraneous stimuli				
9. Is forgetful in daily activities				
10. Fidgets with hands or feet or squirms in seat				
11. Leaves seat in classroom or in other situations in which remaining seated is expected				
12. Runs about or climbs excessively in situations in which it is inappropriate				
13. Has difficulty playing or engaging in leisure activities quietly				
14. Is "on the go" or often acts as if "driven by a motor"				
15. Talks excessively				
16. Blurts out answers before questions have been completed				
17. Has difficulty awaiting turn				
18. Interrupts or intrudes on others (e.g., butts into conversations/games)				
19. Loses temper				
20. Argues with adults				
21. Actively defies or refuses adult requests or rules				
22. Deliberately does things that annoy other people				
23. Blames others for his or her mistakes or misbehavior				
24. Is touchy or easily annoyed by others				
25 Is angry and resentful				
26. Is spiteful or vindictive				

Fig. 1 The MTA SNAP-IV Rating Scale is a revision of the Swanson, Nolan, and Pelham (SNAP) Questionnaire (Swanson, 1992; Swanson et al., 1983). Other versions of the SNAP have been used (see Swanson et al., 1999). For example, the word "Often" has been added to start each item (to maintain consistency with the *DSM-IV* wording), and the rating category "Quite a Bit" has been substituted for "Pretty Much" (as a grammatical concession). Norms are provided by Gaub and Carlson (1997) and Swanson (1992). The items from the *DSM-IV* criteria for attention-deficit/hyperactivity disorder (ADHD) are included for the two subsets of symptoms: Inattention (items 1–9) and Hyperactivity/Impulsivity (items 10–18). Also, items from the *DSM-IV* criteria for oppositional defiant disorder (ODD) are included (items 19–26) because ODD is often present in children with ADHD. Subscale scores on the SNAP-IV (average rating-per-item for Inattention, Hyperactivity/Impulsivity, and Opposition/Defiance) are calculated by summing the scores on the items in the subset and dividing by the number of items in the subset. Additional information on the SNAP-IV Rating Scale is available on the World Wide Web at *ADHD.net*. Permission for clinical use of the SNAP is granted by J.M.S.

ent and teacher sources, subscale scores are calculated by averaging the item scores within the domains of Inattention (Inatt), Hyperactivity/Impulsivity (H/Imp), and Opposition/Defiance (O/D). This generates the six SNAP-IV scores that cover three domains and two sources and that were used in the primary analyses of the MTA (MTA Cooperative Group, 1999a,b).

Scores from rating scales have been described as "soft data" (Kraemer, 1992) because they are affected by variation from the source completing the ratings, as well as the behavior of the individual being evaluated. In general, the correlation of ratings across the parent and teacher sources is relatively low (Achenbach et al., 1987), and this was the case ($r = 0.3$) for the SNAP-IV ratings of the subjects in the MTA (see Swanson et al., 1999). This source variance reduces the precision of measurement of the child's behavior, which attenuates the statistical significance of any treatment effect that might be present (Brown, 1910; Kraemer and Thiemann, 1989; Spearman, 1910). To increase the precision of measurement for the SNAP-IV outcome measures of the MTA study, a summary rating was formed by averaging items from the three domains and the two sources to obtain an overall SNAP-IV_{PT} score. This summary provides a narrowly focused assessment of symptom severity, which was the basis for both the continuous (quantitative) and the categorical (qualitative) outcome measures reported here.

Orthogonal and Paired Comparisons. Basic statistical principles dictate that for four treatment conditions, only three orthogonal (i.e., statistically independent) comparisons are possible. Here we redefine the three main study questions as three post hoc but orthogonal comparisons to help clarify the primary findings of the MTA based on six nonorthogonal paired comparisons.

The first comparison uses all four groups to contrast the average outcome for the two treatments that included the MTA medication algorithm (Comb + MedMgt) and the average of the other two groups that did not (Beh + CC). This will be called the *MTA medication algorithm* comparison ("Comb + MedMgt versus Beh + CC"). The second comparison contrasts the multimodal treatment with one of the unimodal treatments (pharmacological alone). This addresses a major hypothesis of the MTA (as the name suggests), so it will be called the *multimodality superiority* comparison ("Comb versus MedMgt"). The third comparison contrasted the other unimodal treatment (psychosocial alone) with the usual treatments for ADHD in the community that during the years of this study commonly included the use of stimulant medications (MTA Cooperative Group, 1999a,b). This will be called the *psychosocial substitution* comparison ("Beh versus CC"), since families were asked to substitute an uncommonly intensive psychosocial intervention (Taylor, 1999) for the standard treatments for ADHD.

Continuous and Categorical Outcome Measures. For analysis of continuous measures, averages are typically used to summarize group outcomes. The averaging process includes those subjects who had excellent responses (successes) and those who did not (failures), so the group average may not be representative of any individual subject. A categorical outcome measure based on success at the end of treatment provides information (i.e., the percentage of patients with an excellent response to each treatment) that should appeal more to clinical intuition than information from a continuous outcome measure based on group averages of improvement over time.

For this report, an excellent response (and thus "success") was operationally defined by a cutoff of ≤ 1.0 on the SNAP-IV_{PT} score obtained at the end of treatment. This was a logical cutoff based on the *DSM-IV* criteria, which state that low severity of the specified behaviors (i.e., in the range from "not at all" to "just a little" in the subjective ratings on the SNAP-IV) would not be sufficient to qualify as symptoms of ADHD or ODD. This logical cutoff was also consistent with the recommendation of Kazdin and Wilson (1978), who suggested that clinical success should be characterized by the elimination of the presenting problem or loss of symptoms. In addition, norms for the SNAP show that most school-age children do not manifest any of the psychopathology described by the *DSM-IV* items and thus on the SNAP-IV items they receive ratings ≤ 1 (Gaub and Carlson, 1997; Swanson, 1992).

Effect Size and Statistical Significance

The statistically *non*-significant group differences ("Comb versus MedMgt" and "Beh versus CC") are the main sources of controversy about the primary findings of the MTA study. Instead of accepting the null hypothesis about these two comparisons, an evaluation of effect size (ES) can be used to help understand the lack of statistical significance and possible clinical relevance of the findings. In the primary analyses, ES estimates were not explicitly presented but can be derived from the group means and standard deviations (SDs) provided in the initial report (MTA Cooperative Group, 1999a, Table 4). These statistics were based on the available number of subjects for each of the 19 measures obtained at the end of treatment, so missing data may affect the accuracy of the ES estimates, which by our calculations varied from –0.18 to 0.27, with an average ES = 0.05. Here we provide a more thorough evaluation of the ES for both the continuous and categorical measures for this report.

For the continuous measure, the most common ES is the standardized mean difference between groups (Cohen delta, or $d = [M_1 - M_2]/SD_{pooled}$). The mean difference is expressed in terms of SD units, not standard error units (SE = SD/sqrt[N]), so the expected value of ES *does not* increase with sample size. But statistical power (the chance of detecting a true difference) does, so it is often stated that any nonzero ES (even a clinically insignificant effect) can be found statistically significant with a large enough sample size. Therefore, guidelines have been proposed (Cohen, 1988) for classifying ES values as small (0.2), moderate (0.5), and large (0.8) to help interpret the clinical relevance of statistically significant effects. For the present analyses, another relationship is important: ES *does* vary with precision of measurement. It is generally appreciated that an increase in the denominator of the formula (i.e., SD) leads to a decrease in ES, and that imprecise measurement can make the estimated ES smaller and reduce statistical significance. But the reverse is also true: a decrease in SD leads to an increase in ES, so precise measurement can make the estimated ES larger and increase statistical significance. Moreover, a decrease in SD resulting from increased precision of measurement means that the resulting ES better indicates the truth of the matter.

We show here that even procedures implemented after data collection can increase precision and produce a change in the estimated ES. We anticipated that the combination of the six separate SNAP-IV scores into a single summary measure of symptom severity (SNAP-IV_{PT}) would reduce random variation and thus should increase the precision of measurement (Kraemer and Thiemann, 1989). This could increase the estimated ES and affect the statistical significance of the MTA results.

For the categorical measure, the chance of success (p) divided by the chance of failure (1 – p) for each group can be used to estimate of the odds of success (p/[1 – p]). The ratio for two groups is the odds ratio (OR), with a range of zero to infinity. If the groups do not differ, the expected OR = 1. The logarithm of the odds ratio (LOR) provides a range of values (zero now indicating no difference and positive and negative values possible) that is more interpretable than the range of values for OR, and the LOR is used as the ES estimate for comparing group success rates. For normally distributed variables and nonrare outcomes, the relationship between these two measures of ES (Cohen d and LOR) is approximated by $d = LOR/1.6$.

RESULTS

Continuous Outcome Measures

Descriptive statistics were calculated based on end-of-treatment scores. We used the last observation carried forward (LOCF) technique to replace missing observations at the 14-month assessment. We present information in Table 1 for the three domains from parent SNAP-IV ratings ($Inatt_P$, H/Imp_P, and O/D_P) and the three domains from teacher SNAP-IV ratings ($Inatt_T$, H/Imp_T, and O/D_T). Similar descriptive statistics were presented in the primary analyses (MTA Cooperative Group, 1999a, Tables 4 and 5). In the table here, we also present the within-source ($SNAP\text{-}IV_P$ and $SNAP\text{-}IV_T$) and the overall ($SNAP\text{-}IV_{PT}$) summary scores. In addition to the group means (M) in section A, the group SDs are presented in section B, and the ES (Cohen *d*) for the three orthogonal contrasts are presented in section C.

Section B of Table 1 reveals that the SD was reduced substantially for the summary SNAP-IV scores compared to the separate source × domain scores used in the primary analyses. For example, consider the SD values for the Comb group that are presented in the first row of section B. For the $SNAP\text{-}IV_{PT}$ summary, the SD presented in column 12 (0.48) is about 30% less than the average SD (0.69) for the separate domain scores shown in columns 1 to 3 for the parent ratings and in columns 5 to 7 for the teacher ratings ([0.67 + 0.64 + 0.67 + 0.77 + 0.67 + 0.70]/6 = 0.69). This reduction in SD produces a corresponding increase in ES (see section C of the table).

For the MTA medication algorithm contrast, the ES were large enough for each of the six separate SNAP-IV measures (within sources and domains) to reach statistical significance. This confirms and extends the significant medication effects reported in the primary analyses, and it also provides an estimate of the size of the effects: across the six separate SNAP-IV measures the average ES was 0.38, and for the overall summary measure of $SNAP\text{-}IV_{PT}$ the ES increased to 0.59.

For the multimodality superiority comparison, the ES for six separate SNAP-IV measures ranged from $d = 0.11$ to 0.27 (see Table 1, section C, columns 1–3 and 5–7), with an average of 0.19. Two of the six comparisons barely reached the cutoff for statistical significance at $p < .05$. Despite the differences in statistical methods (i.e., a LOCF end-of-treatment analysis versus an ITT–RR analysis), these results differ only slightly from those reported for the primary analyses (MTA Cooperative Group, 1999a), which indicated that outcome tended to be better in the Comb than the MedMgt condition, but that the difference was not quite statistically significant for any domain. In addition to confirming the primary analyses, these secondary analyses also show the impact of increasing precision of measurement: for the overall summary score ($SNAP\text{-}IV_{PT}$), the estimated ES was still small to moderate (0.26) but was now statistically significant at $p < .05$.

The secondary analysis also confirms the primary finding that the psychosocial substitution effects were small and statistically nonsignificant. The ES for the individual SNAP-IV measures varied from 0.00 to 0.18 (with an average of 0.09), and the ES for the summary $SNAP\text{-}IV_{PT}$ measure was also nonsignificant and remained as small (0.09).

Categorical Outcome Measures

The ordering of the treatment groups on the categorical outcome measure of successful treatment mirrored the ordering based on the continuous outcome measure of symptom severity. The success rate was highest for the Comb treatment group (68%), followed by the MedMgt group (56%), the Beh group (34%), and the CC group (25%). To evaluate group differences on this categorical outcome measure, we used a logistic regression analysis in which the independent variables were site (6 levels, with 5 *df*), treatment (4 levels, with 3 *df*), and site × treatment ($6 \times 4 = 24$ levels, with $5 \times 3 = 15$ *df*). To separate the treatment effect into nonoverlapping components, the 3 *df* were used to evaluate the previously described orthogonal contrasts. To separate the site × treatment interaction effects into nonoverlapping components, the 15 *df* were apportioned to the 3 orthogonal contrasts, each with 5 *df*. Table 2 shows that the medication algorithm contrast ("Comb + MedMgt versus Beh + CC") was highly significant ($\chi^2 = 58.8$, $df = 1$, $p < .001$) and that the multimodality superiority contrast ("Comb versus Med") was just significant at conventional levels ($\chi^2 = 4.5$, $df = 1$, $p < .04$). The psychosocial substitution contrast ("Beh versus CC") was not statistically significant at conventional levels ($\chi^2 = 2.7$, $df = 1$, $p = .10$) and was so small that even if made statistically significant (by increasing precision of measurement or sample size), it would have no clinical significance.

As recommended by Kraemer (1992), we present in Figure 2 the cumulative distribution curves of the $SNAP\text{-}IV_{PT}$ scores to emphasize the clinical relevance and practical significance of MTA treatment group differences. These curves present the percentage of cases in each group with $SNAP\text{-}IV_{PT}$ scores better than specific cutoff values across

TABLE 1
Summary Statistics for the SNAP-IV Rating Scale

	Parent				Teacher				Parent/Teacher Average			
Column No.	1	2	3	4	5	6	7	8	9	10	11	12
Domain	$Inatt_P$	H/Imp_P	O/D_P	**$SNAP\text{-}IV_P$**	$Inatt_T$	H/Imp_T	O/D_T	**$SNAP\text{-}IV_T$**	$Inatt_{PT}$	H/Imp_{PT}	O/D_{PT}	**$SNAP\text{-}IV_{PT}$**
Section A: mean ratings[a]												
Comb	1.08	0.88	0.80	**0.92**	1.16	0.76	0.63	**0.85**	1.11	0.82	0.72	**0.88**
MedMgt	1.18	0.96	0.99	**1.04**	1.28	0.96	0.80	**1.00**	1.23	0.96	0.90	**1.02**
Beh	1.40	1.21	1.05	**1.22**	1.51	1.16	1.00	**1.23**	1.45	1.18	1.02	**1.23**
CC	1.49	1.34	1.11	**1.32**	1.47	1.21	1.01	**1.22**	1.48	1.28	1.06	**1.27**
Section B: SD of ratings[b]												
Comb	0.67	0.64	0.67	**0.57**	0.77	0.67	0.70	**0.58**	0.61	0.53	0.56	**0.48**
MedMgt	0.76	0.69	0.77	**0.65**	0.84	0.80	0.81	**0.72**	0.65	0.62	0.65	**0.57**
Beh	0.67	0.70	0.74	**0.60**	0.80	0.81	0.82	**0.70**	0.59	0.63	0.62	**0.53**
CC	0.70	0.71	0.66	**0.58**	0.83	0.82	0.83	**0.66**	0.62	0.55	0.54	**0.45**
Section C: ES[c]												
Medication algorithm	0.45***	0.52***	0.25*	**0.48*** **	0.33**	0.42***	0.36**	**0.45*** **	0.48***	0.58***	0.39**	**0.59*** **
Multimodal superiority	0.14	0.11	0.26*	**0.20**	0.15	0.27*	0.23	**0.23**	0.18	0.25*	0.29*	**0.26***
Psychosocial substitution	0.13	0.18	0.08	**0.16**	0.05	0.07	0.00	**0.01**	0.04	0.16	0.05	**0.09**

Note: SNAP-IV = Swanson, Nolan, and Pelham, version IV; Inatt = Inattention; H/Imp = Hyperactivity/Impulsivity; O/D = Opposition/Defiance; Comb = combined treatment; MedMgt = medication management; Beh = behavioral treatment; CC = community comparison.

[a] The mean values for each treatment group across three domains (Inatt, H/Imp, O/D), for the composite across domains within the two sources ($SNAP\text{-}IV_P$ and $SNAP\text{-}IV_T$), and for the composite across sources and domains ($SNAP\text{-}IV_{PT}$).

[b] The standard deviation values for each mean value.

[c] The effect size (ES) estimated by the Cohen *d* for three orthogonal comparisons of the treatment groups (medication algorithm [Comb + MedMgt vs. Beh + CC], multimodality superiority [Comb vs. MedMgt], and psychosocial substitution [Beh vs. CC]), with statistical significance level shown (* $p < .05$; ** $p < .01$; *** $p < .001$).

TABLE 2
Effect Size Expressed as LOR for Logistic Regression Analysis of the Qualitative Outcome Measure, for the Main Effects Defined by Orthogonal Contrasts and Six Sites, Plus the Interactions of Contrast by Site

	(Comb+MedMgt vs. Beh+CC)	(Comb vs. MedMgt)	(Beh vs. CC)
Main effect (contrast)			
χ^2, *df* = 1	58.8	4.5	2.7
p value	.00	.04	.10
LOR	2.79	0.53	0.43
Interaction with site			
χ^2, *df* = 5	3.8	6.7	14.9
p value	.58	.25	.01
LOR by site			
1 (Pittsburgh)	1.92	0.42	–1.00
2 (Irvine)	3.82	0.00	–0.24
3 (New York City)	3.07	1.99	0.61
4 (Berkeley)	2.44	0.43	–0.35
5 (Durham)	2.21	0.00	1.85
6 (Queens/Montreal)	3.70	0.29	1.55

Note: LOR = logarithm of the odds ratio; Comb = combined treatment; MedMgt = medication management; Beh = behavioral treatment; CC = community comparison.

the possible range of values of the outcome measure (i.e., from 0 to 3, specified in Fig. 2 at convenient increments of 0.25). The logical cutoff that we selected to denote successful treatment (SNAP-$IV_{PT} \leq 1$) is shown. In the top panel of Figure 2, we present the distribution of this narrow summary score for the four MTA treatment groups before time 0 (at phase B of the assessment: Hinshaw et al., 1997) and for a local normative control group of 288 randomly selected students from the same schools. This graph shows that most students (88%) in the school population receive SNAP-IV_{PT} ratings below our cutoff for symptom presence (≤1.0) and that (due to the selection criteria) all MTA subjects had ratings above the cutoff. In the bottom panel of Figure 2, we present the distributions of SNAP-IV_{PT} scores for the MTA treatment groups after 14 months of treatment. The percentages of patients meeting our operational definition of success at the end of treatment are shown for the four treatment groups (Comb = 68%, MedMgt = 56%, Beh = 34%, and CC = 25%).

In the logistic regression analysis, the site × treatment interactions were not significant for the medication algorithm comparison ($\chi^2 = 3.8$, $p = .58$) or the multimodality superiority comparison ($\chi^2 = 6.7$, $p = .25$). However, the site × treatment interaction was significant for the psychosocial substitution comparison ($\chi^2 = 14.9$, $df = 5$, $p < .01$), which was the only comparison that did not produce a significant main effect. This surprising result may provide some clarification of one of the primary findings (i.e., that the intensive psychosocial treatment provided by the MTA was not superior to the community care treatments).

The ES estimates (LORs) for each site are presented in Table 2. The site differences for this contrast may be best viewed in receiver operating characteristic (ROC) curves, which present the cumulative percentages of success for the Beh group (on the ordinate) relative to the percentage of success for the CC group (on the abscissa). (For exploratory purposes, the ROC curves for the Comb and MedMgt groups relative to the CC group are also shown in Fig. 3. Even though these two-group contrasts were not part of the orthogonal comparison method used in this report, they were evaluated in the part of the primary analyses that compared all MTA treatments to the community control.) An upward "bow" in the ROC curve indicates a higher success rate than the CC group, while a downward "bow" indicates a lower success rate. The ROC curves in Figure 3 and the LOR values in Table 2 reveal that the Beh treatment was better than the CC treatment at site 3 (New York City), site 5 (Queens/Montreal), and site 6 (Durham), but the reverse was true for site 1 (Pittsburgh), site 2 (Irvine), and site 4 (Berkeley).

Even though the site × treatment interaction was not statistically significant for the multimodality superiority comparison, an exploratory inspection of Figure 3 and Table 2 suggests that site differences contributed to the lack of significance. As shown in Table 2, at site 3 (New York City) the combination of intensive psychosocial treatment added a large boost to the effect of pharmaco-

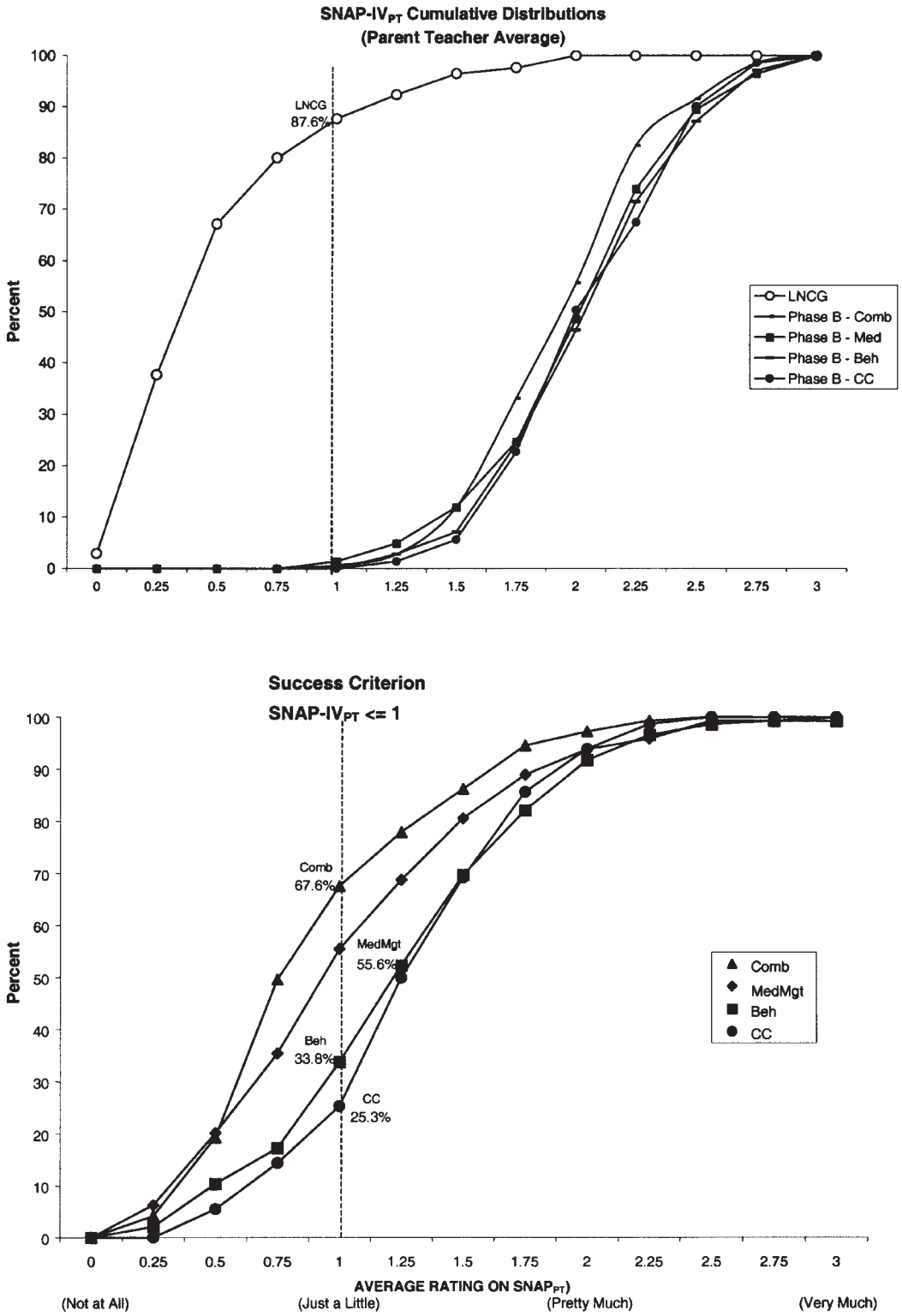

Fig. 2 Top panel: cumulative distribution curves for combined treatment (Comb), medication management (MedMgt), behavioral treatment (Beh), and community comparison (CC) groups and for the local normative control group (LNCG), showing percentage with Swanson, Nolan, and Pelham, version IV, parent and teacher (SNAP-IV$_{PT}$) ratings ≤ 1 before entering the MTA. Bottom panel: cumulative distribution curves for the Comb, MedMgt, Beh, and CC groups showing percentage with SNAP-IV$_{PT}$ ratings ≤ 1 after 14 months of treatment.

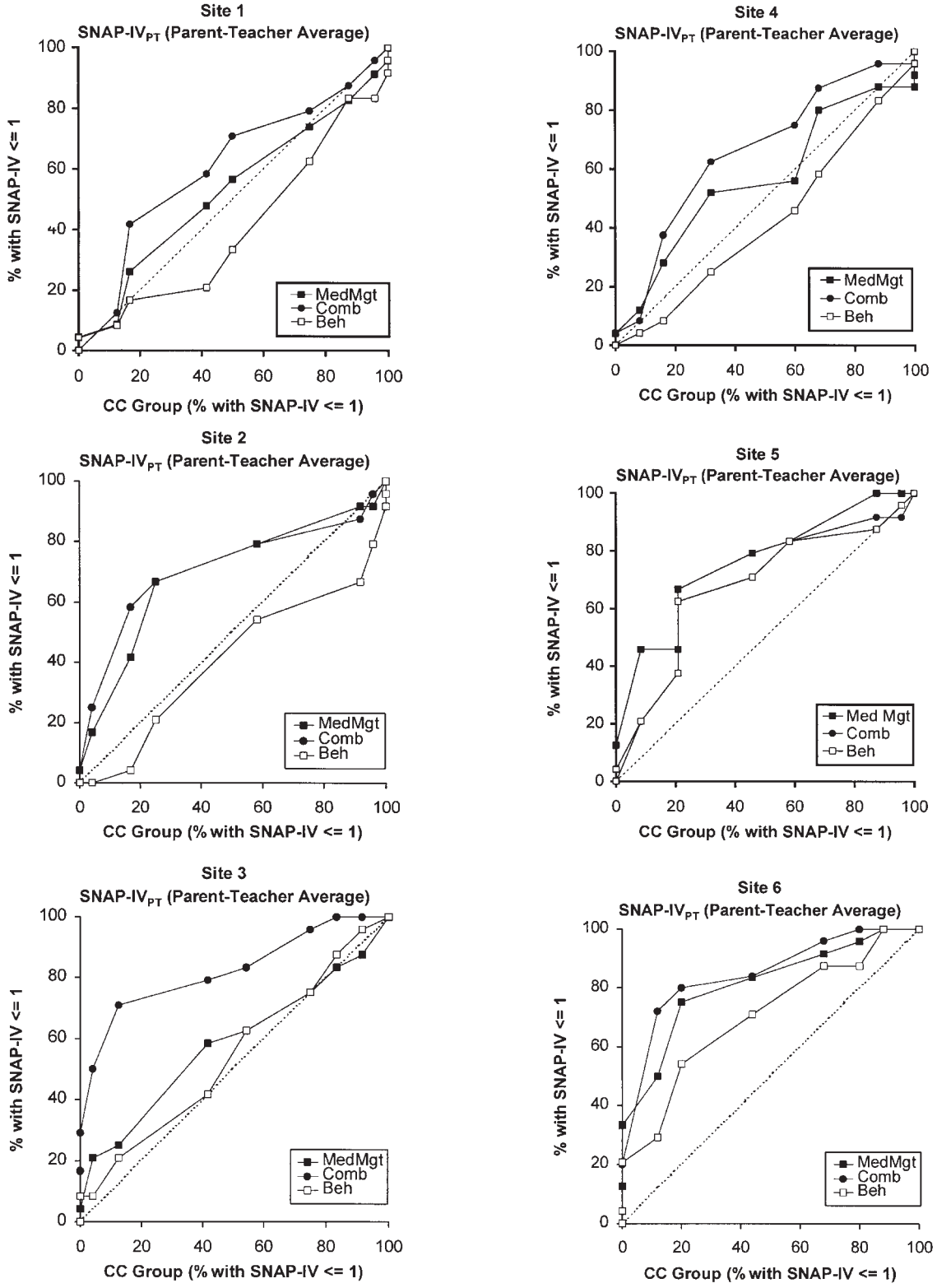

Fig. 3 Receiver operating characteristic curves for the comparison of the MTA treatment groups (MedMgt, Beh, and Comb) to the CC group, separately for the six sites of the MTA. SNAP-IV$_{PT}$ = Swanson, Nolan, and Pelham, version IV, parent and teacher ratings; MedMgt = medication management; Comb = combined treatment; Beh = behavioral treatment; CC = community comparison.

logical treatment alone (LOR = 1.99) and was positive at 3 other sites: LORs = 0.42 at site 1(Pittsburgh), 0.43 at site 4 (Berkeley), and 0.29 at site 6 (Queens/Montreal). However, the LORs were 0 (negligible) at site 2 (Irvine) and site 5 (Durham), which were the sites with major responsibilities for developing and monitoring the paraprofessional and parent training components of the psychosocial treatment.

DISCUSSION

Three Clarifications of the Primary Findings

The secondary analyses presented here confirm and extend the results of the primary analyses (MTA Cooperative Group, 1999a,b). First, these secondary analyses confirm the large ES (Cohen *d* = 0.59; LOR = 2.79) associated with exposure to the MTA Medication Algorithm ("Comb + MedMgt versus Beh + CC"). The logistic regression analysis revealed that the ES were consistently large at all sites in this study. The clinical impact on "success rate" was substantial: under conditions of the MTA, exposure to the medication algorithm in the Comb and MedMgt groups resulted in more than twice as many patients reaching the stringent criterion of successful treatment (average success rate = 62%) as nonexposure in the Beh and CC conditions (average success rate = 30%).

Second, these secondary analyses confirm that the magnitude of the multimodality superiority effect was small to moderate and was variable across domains, sources, and sites. When random variability was reduced by averaging the SNAP-IV outcome measures across sources and domains, the estimated ES (0.26) was now statistically significant, but this incremental improvement with multimodality treatment would still be considered small by conventional standards (Cohen, 1988). The success rates presented here may be more useful to clinicians faced with making decisions about the use of these two treatments. The multimodal treatment produced about 12% more successes than the unimodal treatment with medication alone (68% versus 56%), which represented a 21.4% increase in the rate of excellent response. These perspectives, based on a categorical outcome measure, may help clinicians judge the clinical relevance of the small multimodality superiority effect suggested by the primary analyses presented earlier (MTA Cooperative Group, 1999a,b) and elucidated by the secondary analyses presented here.

Third, these secondary analyses suggest that the inability to demonstrate statistically or clinically significant effects of intensive psychosocial intervention over treatment as usual ("Beh versus CC") may be due to local conditions. In the communities represented in the MTA, moderate to large positive effects occurred at three sites and small to large negative effects occurred at the other three sites. Thus, even though the overall psychosocial substitution effect was negligible, this average effect did not describe the outcome at any individual site. It is possible (and likely) that the overall lack of a psychosocial substitution effect may depend on the frequency and quality of treatments obtained in the community (which were not controlled) rather than the impact of the standardized and monitored behavioral interventions provided by the protocol of the MTA study. The variability in local conditions may depend on differences in the demographics of the patients across sites (e.g., in socioeconomic status), in what treatment as usual might mean in the different communities (e.g., in the use of medication or in the available psychosocial interventions), in the implementation of the psychosocial protocol (e.g., high or low cooperation in school settings), or in a variety of other uncontrolled factors.

Limitations

This report has several limitations. Some are design features shared with the primary outcome report, such as (1) the decision to fade the psychosocial intervention in the Beh and Comb treatments and to continue the pharmacological intervention in the MedMgt and Comb treatments, which mimics clinical practice but may bias the outcome in favor of the effects of medication; and (2) the lack of a placebo-treated or untreated control group, which may distort (reduce) the apparent size of the impact of Beh treatment compared with the CC treatment, which in most cases included an active treatment with stimulant medication. These limitations have been discussed in detail elsewhere (National Institutes of Health, 2000; Pelham, 1999; Taylor, 1999).

Other potential limitations are unique to the secondary analyses reported here. One is the use of a narrow outcome measure, defined by the symptoms of ADHD and ODD and evaluated by a single instrument (the SNAP-IV rating scale). This emphasis is based on a narrow view of psychiatric disorder and is consistent with a basic premise of the *DSM-IV* and *ICD-10* criteria for ADHD and ODD (i.e., symptom presence that is chronic over

time and pervasive across settings) and with the power analyses that focused on core ADHD symptoms to set the sample size for the MTA. However, intensity of symptoms certainly varies over time, and the level might not be the same at different assessment points.

To address this limitation, Conners et al. (2001) developed a composite from a large number of the outcome measures derived from multiple instruments of the assessment battery. Different views of psychiatric disorder were adopted and different procedures were used to develop this broad composite than for our narrow summary based on *DSM-IV* symptoms. The broad composite was designed to capture expected fluctuation in behavior over time and across settings, rather than a stable and chronic pattern of behavior. This broad composite differed in many ways from the narrow summary used here. Separate scores from many instruments (instead of one) and many domains (instead of just six) were used as a starting point, and a *z*-transform was used to normalize scores. A factor analysis of the MTA baseline data was performed to identify domains (instead of relying on the *DSM* framework). The factor scores were averaged so the sources would contribute equally (instead of using instruments with the same items for both sources). In the analysis of the broad composite, multiple comparisons were used (rather than orthogonal comparisons). Despite these differences, the main clarifications about the nonsignificant effects of the primary analyses were very similar to the clarifications based on the narrow outcome measure. For the broad measure, the previously nonsignificant multimodality superiority effect was increased and rendered significant at $p < .05$ (with an ES = 0.28, about the same as reported here), but the psychosocial substitution effect remained negligible and nonsignificant (ES = 0.09, exactly the same as reported here). However, these two approaches did produce some differences in the magnitude of the significant effects reported in the primary analyses. Conners et al. (2001) present ES estimates for the six nonorthogonal comparisons, which include two relevant to the significant effects in the primary analyses: for MedMgt versus CC, $d = 0.35$ and for Comb versus CC, $d = 0.70$. We calculated ES for these two comparisons based on narrow SNAP-IV_{PT} outcome measures (for MedMgt versus CC, $d = 0.45$ and for Comb versus CC, $d = 0.83$), which were larger than the ES reported by Conners et al. (2001) for the broad measure. This suggests that the medication effects on noncore symptoms of ADHD (included in the broad composite based on multiple instruments) may be smaller than on core symptoms of ADHD and ODD (the sole components of the narrow summary based on the SNAP-IV).

The use of a logical cutoff (SNAP-$IV_{PT} \leq 1$) may be considered a limitation by some. Other operational definitions of "success" could be used, such as the statistical procedure proposed by Jacobson and Truax (1991), who operationally defined a statistical cutoff for "normalization" as the midpoint between the clinical and control groups. Even though we have methodological concerns because this relies on between-subject variability to make decisions about individual subjects, we applied the Jacobson and Truax method using norms for a control group (derived from SNAP-IV ratings of 288 classmates of the MTA subjects) and a clinical group (derived from the baseline SNAP-IV assessments of the MTA subjects). The resulting statistical cutoff was a SNAP-IV_{PT} rating ≤ 1.2, and this was less conservative than the logical cutoff (SNAP-IV_{PT} rating ≤ 1.0) used here. However, the cumulative distribution curves are approximately parallel for values surrounding 1.0 (see Fig. 2), so the use of the Jacobson and Truax cutoff value of 1.2 results in the same ordering and about the same relative positions of the four treatment groups in terms of success rates (i.e., Comb = 75%, MedMgt = 63%, Beh = 46%, and CC = 42%) as the use of our cutoff of 1.0.

Clinical Implications

The clinical relevance of the primary results can be clarified by addressing three questions that groups of clinicians (who are engaged in "treatment as usual") may ask about the MTA procedures and results. First, If the MTA medication algorithm is adopted, how many more patients would show a successful response to treatment? The secondary analyses suggest that the number of cases successfully treated would more than double (from 30% to 62%). Second, If the MTA psychosocial algorithm is combined with the medication algorithm, how many patients would achieve a successful response? The secondary analyses suggest that 12% more of the total number of patients (an increase from 56% to 68%) would meet the criteria for an excellent response, which represents an increase of more than 20% in success rate ([68% – 56%]/56% = 21.4%). Third, If the use of medication is rejected as an option and all patients were treated with the MTA psychosocial algorithm instead, what would happen to the success rate? The secondary analyses suggest that the impact of such a drastic change could be positive or negative, depending on local conditions and characteristics of the usual practices.

The NIMH Collaborative Multisite Multimodal Treatment Study of Children With Attention-Deficit/Hyperactivity Disorder (MTA) is a cooperative treatment study performed by six independent research teams in collaboration with the staff of the Division of Services and Intervention Research of the NIMH, Rockville, MD, and the Office of Special Education Programs (OSEP) of the U.S. Department of Education (DOE). The NIMH Principal Collaborators are Peter S. Jensen, M.D., L. Eugene Arnold, M.Ed., M.D., John E. Richters, Ph.D., Joanne B. Severe, M.S., Donald Vereen, M.D., and Benedetto Vitiello, M.D. Principal Investigators and Coinvestigators from the six sites are as follows: University of California at Berkeley/San Francisco (UO1 MH50461): Stephen P. Hinshaw, Ph.D., Glen R. Elliott, M.D., Ph.D.; Duke University (UO1 MH50447): C. Keith Conners, Ph.D., Karen C. Wells, Ph.D., John S. March, M.D., M.P.H.; University of California at Irvine/Los Angeles (UO1 MH50440): James M. Swanson, Ph.D., Dennis P. Cantwell, M.D., Timothy Wigal, Ph.D.; Long Island Jewish Medical Center/Montreal Children's Hospital (UO1 MH50453): Howard B. Abikoff, Ph.D., Lily Hechtman, M.D.; New York State Psychiatric Institute/Columbia University/Mount Sinai Medical Center (UO1 MH50454): Laurence L. Greenhill, M.D., Jeffrey H. Newcorn, M.D.; University of Pittsburgh (UO1 MH50467): William E. Pelham, Ph.D., Betsy Hoza, Ph.D. Helena C. Kraemer, Ph.D. (Stanford University) is statistical and design consultant. The OSEP/DOE Principal Collaborator is Ellen Schiller, Ph.D.

REFERENCES

Achenbach TM, McConaughy SH, Howell CT (1987), Child/adolescent behavioral and emotional problems: implications of cross-informant correlations for situational specificity. *Psychol Bull* 10:213–232

American Psychiatric Association (1994), *Diagnostic and Statistical Manual of Mental Disorders, 4th edition (DSM-IV).* Washington, DC: American Psychiatric Association

Brown W (1910), Some experimental results in the correlation of mental abilities. *Br J Psychol* 3:296–322

Cohen J (1988), *Statistical Power Analysis for the Behavioral Sciences,* 2nd ed. Hillsdale, NJ: Erlbaum

Conners CK, Epstein JN, March JS et al. (2001), Multimodal treatment of ADHD in the MTA: an alternative outcome analysis. *J Am Acad Child Adolesc Psychiatry* 40:159-167

Gaub M, Carlson C (1997), Behavioral characteristics of *DSM-IV* ADHD subtypes in a school-based population. *J Abnorm Child Psychol* 25:103–111

Hinshaw SP, March JS, Abikoff H et al. (1997), Comprehensive assessment of childhood attention-deficit hyperactivity disorder in the context of a multisite, multimodal clinical trial. *J Attention Disord* 1:217–234

Jacobson NS, Truax P (1991), Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *J Consult Clin Psychol* 59:12–19

Kazdin AE, Wilson GT (1978), *Evaluation of Behavior Therapy: Issues, Evidence and Research Strategies*. Cambridge, MA: Ballinger

Kraemer HC (1992), Reporting the size of effects in research studies to facilitate assessment of practical or clinical significance. *Psychoneuroendocrinology* 17:527–536

Kraemer HC, Thiemann S (1989), A strategy to use soft data effectively in randomized controlled clinical trials. *J Consult Clin Psychol* 57:148–154

MTA Cooperative Group (1999a), 14-month randomized clinical trial of treatment strategies for attention deficit hyperactivity disorder. *Arch Gen Psychiatry* 56:1073–1086

MTA Cooperative Group (1999b), Effects of comorbid anxiety disorder, family poverty, session attendance, and community medication on treatment outcome for attention-deficit hyperactivity disorder. *Arch Gen Psychiatry* 56:1088–1096

National Institutes of Health (2000), National Institutes of Health Consensus Development Conference Statement: Diagnosis and Treatment of Attention-Deficit/Hyperactivity Disorder (ADHD). *J Am Acad Child Adolesc Psychiatry* 39:182-193

Pelham WE (1999), President's message: the NIMH Multimodal Treatment Study for ADHD: just say yes to drugs. *Clin Child Psychol Newsl* 14:1–6

Spearman C (1910), Correlation calculated from faulty data. *Br J Psychol* 3:271–295

Swanson JM (1992), *School-Based Assessments and Interventions for ADD Students.* Irvine, CA: KC Publishing

Swanson JM, Lerner MA, March J, Gresham FM (1999), Assessment and intervention for attention-deficit/hyperactivity disorder in the schools: lessons from the MTA study. *Pediatr Clin North Am* 46:993–1009

Swanson JM, Sandman CA, Deutsch C, Baren M (1983), Methylphenidate (Ritalin) given with or before breakfast, Part I: behavioral, cognitive, and electrophysiological effects. *Pediatrics* 72:49–55

Swanson JM, Sergeant JA, Taylor E, Sonuga-Barke EJS, Jensen PS, Cantwell DP (1998), Attention-deficit hyperactivity disorder and hyperkinetic disorder. *Lancet* 351:429–433

Taylor E (1999), Development of clinical services for attention-deficit/hyperactivity disorder. *Arch Gen Psychiatry* 56:1097–1099

World Health Organization (1992), *ICD-10 Classification of Mental and Behavioral Disorders: Clinical Descriptions and Diagnostic Guidelines.* Geneva: World Health Organization

World Health Organization (1993), *ICD-10 Classification of Mental and Behavioral Disorders: Diagnostic Criteria for Research.* Geneva: World Health Organization