Consistency of Atypical Antipsychotic Superiority to Placebo in Recent Clinical Trials

Scott W. Woods, Marilyn Stolar, Michael J. Sernyak, and Dennis S. Charney

Background: The use of control placebos in clinical trials of new antipsychotic medications is increasingly under examination. The active controlled equivalence study could offer a potential alternative design. First, however, it must be clear that any proposed standard control agent has been consistently superior to placebo in previous studies.

Methods: Through a Freedom of Information Act request, we identified nine placebo-controlled trials of risperidone, olanzapine, or quetiapine.

Results: Meta-analysis indicated that the pooled estimate of the true population effect size ± SE was 0.46 ± 0.06 for categorical response rates and >0.53 ± 0.07 for the continuous Brief Psychiatric Rating Scale change score outcome measure. If the desired detectable effect size is set very conservatively at a 95% confidence lower bound for the estimate of true effect size, statistical power for random samples of 80 per group drawn from a population of subjects similar to that of the nine meta-analyzed studies is .67 for categorical response rates and >.82 for the continuous measure, based on one-sided α = .05.

Conclusions: These data suggest substantial confidence that a therapeutic dose of an atypical antipsychotic will be statistically superior to placebo in an adequately sized randomized trial, when reporting a continuous measure as the principal outcome. Biol Psychiatry 2001;49:64-70 © 2001 Society of Biological Psychiatry

Key Words: Equivalence, atypical antipsychotics, risperidone, olanzapine, quetiapine, meta-analysis

Introduction

The ethics of the continued use of control placebos in clinical trials of new antipsychotic medications has come under increasing scrutiny in recent years (Addington 1995; Carpenter et al 1997; Streiner 1999; Weijer 1999; Zipursky and Darby 1999). One of the strongest arguments supporting the placebo-controlled design in clinical trials for schizophrenia is the view that no other design will satisfy the societal imperative that available treatments be proven effective. U.S. Food and Drug Administration (FDA) policy reflects this view, requiring two or more adequately controlled trials as demonstration of efficacy. Control groups considered adequate to provide evidence of antipsychotic efficacy are limited to placebos or other control treatments that are relatively ineffective, such as a subtherapeutic dose of an active medication.

The active controlled equivalence study, where the new medication is evaluated for equivalence to a currently used standard medication, is a potential alternative design in which every patient receives a treatment that is believed or hoped to be effective. Equivalence should not be confused with failing to detect a significant difference in an underpowered study; rather, establishing equivalence requires setting a threshold difference beyond which the medications would not be considered equivalent and demonstrating that any observed difference is significantly smaller than the threshold. The equivalence design, however, has been criticized on several counts (Gould 1991; Jones et al 1996; Leber 1989, 1991; Pledger and Hall 1986; Senn 1993; Temple 1982, 1983, 1996, 1997; Tramer et al 1998; Weijer 1999). These criticisms include the following: that there is no accepted criterion for the equivalence threshold, that large samples may be required, that without placebos an active control may be biased toward greater apparent effectiveness, and that not seeking to establish a difference between treatments may lead to methodological sloppiness.

Perhaps the most important objection to the equivalence design is that it is not possible to know whether both medications worked or whether neither worked. Simply because the active standard medication is effective in general does not mean it necessarily was or will be found superior to placebo in every trial. Interpretation of an equivalence trial requires making the assumption that the standard treatment comparator was effective in the current study. This activity must be assumed, because the equivalence design provides no internal evidence to support or reject the effectiveness of the standard treatment control within the study. Thus equivalence trials do not have

From the Treatment Research Program, Connecticut Mental Health Center (SWW) and the Department of Psychiatry, Yale University School of Medicine (SWW, MS, MJS, DSC), New Haven, and the West Haven Veterans Affairs Medical Center, West Haven (MS, MJS, DSC), Connecticut.

Address reprint requests to Scott W. Woods, M.D., Yale University School of Medicine, Department of Psychiatry, 34 Park Street, New Haven CT 06519. Received November 23, 1999; revised June 21, 2000; accepted June 27, 2000.

© 2001 Society of Biological Psychiatry
### Methods and Materials

#### Study Identification

We searched for placebo-controlled trials of the three newer atypical antipsychotic medications currently marketed in the United States, risperidone, olanzapine, and quetiapine, in the following sources: MEDLINE, the bibliographies of identified reports, and via a Freedom of Information (FOI) Act request from the FDA for each medication. The data available from the FDA through FOI allows identification of trials unaffected by publication bias. For each study the following information was identified: inpatient versus outpatient status, trial duration, the reported a priori responder definition, response rates, end point change scores on the Brief Psychiatric Rating Scale (BPRS), and the reported statistical significance of each atypical antipsychotic arm versus placebo.

#### Data Preparation

In general, little preparation of the data was required before meta-analysis. Standard deviations for the BPRS change scores were available for all studies except one (risperidone study 0201; Table 1). The maximum possible SE of the difference was calculated for study 0201 from the reported inexact p-value ($p < .001$) and then converted to yield a maximum effect size denominator, which in turn yields a minimum effect size estimate. The SD for the united two separately published portions of another study (risperidone study 0204; Table 1) were calculated exactly from published separate standard deviations.

---

### Table 1. Summary of Methods in Placebo-Controlled Newer Antipsychotic Studies

<table>
<thead>
<tr>
<th>Newer antipsychotic</th>
<th>Studya</th>
<th>Duration</th>
<th>Patientsb</th>
<th>Percent receiving a placebo</th>
<th>Placebo</th>
<th>Reported response definitionc</th>
<th>Adjustedd definition</th>
<th>Placebo response rate ± SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Risperidone</td>
<td>0201b</td>
<td>6 weeks</td>
<td>S, I (1)</td>
<td>33%</td>
<td>53</td>
<td>20% reduction in BPRS$_{0–6}$</td>
<td>20%</td>
<td>.30 ± .06</td>
</tr>
<tr>
<td></td>
<td>0204f</td>
<td>8 weeks</td>
<td>S, I (2.5)</td>
<td>17%</td>
<td>86</td>
<td>20% reduction in PANSS at end point</td>
<td>23%</td>
<td>.20 ± .04</td>
</tr>
<tr>
<td></td>
<td>72g</td>
<td>4 weeks</td>
<td>S, I (2)</td>
<td>33%</td>
<td>79</td>
<td>20% reduction in PANSS at end point</td>
<td>23%</td>
<td>.47 ± .06</td>
</tr>
<tr>
<td>Olanzapine</td>
<td>HGAP</td>
<td>6 weeks</td>
<td>S, I (2.8)</td>
<td>33%</td>
<td>49</td>
<td>40% reduction in BPRS$_{0–6}$ or final</td>
<td>40%</td>
<td>.08 ± .04</td>
</tr>
<tr>
<td></td>
<td>HGAD</td>
<td>6 weeks</td>
<td>S, I (2.8)</td>
<td>20%</td>
<td>62</td>
<td>BPRS ≤ 18, and week 3 completion</td>
<td>40%</td>
<td>.32 ± .06</td>
</tr>
<tr>
<td>Quetiapine</td>
<td>0004</td>
<td>3 weeks</td>
<td>S, I</td>
<td>50%</td>
<td>4</td>
<td>30% reduction in BPRS$_{1–7}$</td>
<td>35%</td>
<td>.50 ± .25</td>
</tr>
<tr>
<td></td>
<td>0006</td>
<td>6 weeks</td>
<td>S, I (6.5)</td>
<td>50%</td>
<td>53</td>
<td>not worsened on CGI at end point</td>
<td>0%</td>
<td>.58 ± .07</td>
</tr>
<tr>
<td></td>
<td>0008</td>
<td>6 weeks</td>
<td>S, I (3.5)</td>
<td>33%</td>
<td>92</td>
<td>30% reduction in BPRS$_{0–6}$ at any time</td>
<td>30%</td>
<td>.37 ± .05</td>
</tr>
<tr>
<td></td>
<td>0013</td>
<td>6 weeks</td>
<td>S, I (4.5)</td>
<td>14%</td>
<td>51</td>
<td>30% reduction in BPRS$_{0–6}$ at any time</td>
<td>30%</td>
<td>.35 ± .05</td>
</tr>
</tbody>
</table>

---

* S, schizophrenia; I, enrolled as inpatients; PANSS, Positive and Negative Syndrome Scale; CGI, Clinical Global Impressions.
* bNumber in parentheses is an estimate of mean required length of hospital stay in weeks based on methods. 0201: required to stay 1 week after enrollment. 0204: required to stay until the end of the dose titration period, roughly 2–3 weeks after enrollment. 72: required to stay 2 weeks after enrollment. HGAP and HGAD: required to stay 2 weeks after randomization, plus 4–9 days placebo run-in. 0004: required stay not reported. 0006: required to stay throughout study, including a minimum of 2 days placebo run-in. 0008: required to stay 3–4 weeks after enrollment. 0201: required to stay 1 week after enrollment. 0204: required to stay 2 weeks after enrollment. HGAP and HGAD: required to stay 2 weeks after randomization, plus 3–7 day placebo run-in.
* cThe definition for each study is the authors’ a priori definition. Brief Psychiatric Rating Scale (BPRS) items can be scored 0–6 or 1–7.
* dEmploying the reported percent reduction in BPRS for studies 0201, 0008, and 0013. Only the 40% reduction component of the complex criteria for olanzapine studies HGAP and HGAD were used. Scoring reported as 1–7 adjusted to 0–6 for studies 0204, 72, and 0004. Percent reduction for study 0006 was set at zero.
* eOnly partial data for 0201 have been published (Borison et al 1992). Full data available from the U.S. Food and Drug Administration via the Freedom of Information (FOI) Act.
* fThe Canadian sites (Chouinard et al 1993) and the U.S. sites (Marder and Meibach 1994) for study 0204 were published separately.
* gThis study is reported in FOI Act supplemental material but has been published only in abstract form (Potkin 1997).

---

Internal “assay sensitivity” (Temple 1997), and internal assay sensitivity is clearly of great value.

On the other hand, the necessity for internal assay sensitivity, rather than the value of it, has often been supported by evidence of historical inconsistency in tests of the efficacy of standard comparators. If a standard comparator could be demonstrated to have been consistently effective compared to a placebo, this historical consistency would provide one basis for a strong argument, albeit not internal proof, that the standard comparator would probably be effective in an equivalence study also.

An example from depression research is frequently cited, in which six studies of a new antidepressant were conducted, each with a design that included the new drug, a control placebo, and an active standard (Leber 1989, 1991). In all six studies the new drug performed similarly to the active standard, but in five of the six studies neither the new drug nor the active standard was significantly superior to the placebo. This example has been used to suggest that an equivalence design could have resulted in an ineffective antidepressant being approved for prescription use.

But what about antipsychotics for schizophrenia? How consistent is the recent historical record of antipsychotic superiority to placebo for schizophrenia? This is the question this article seeks to address.
Table 2. Meta-Analysis of Newer Antipsychotics versus Placebo on Categorical Response Rate

<table>
<thead>
<tr>
<th>Newer antipsychotic</th>
<th>Study(^a)</th>
<th>Therapeutic dose arm(s)(^b)</th>
<th>Therapeutic dose response rate(^c) (n)</th>
<th>Placebo response rate(^c) (n)</th>
<th>Effect size Cohen’s (h) ± SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Risperidone</td>
<td>0201</td>
<td>(7.8 mg)</td>
<td>.65 (51)</td>
<td>.30 (53)</td>
<td>0.71 ± 0.20(^d)</td>
</tr>
<tr>
<td></td>
<td>0204</td>
<td>6 mg</td>
<td>.61 (85)</td>
<td>.20 (86)</td>
<td>0.87 ± 0.15(^d)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10 mg</td>
<td>.39 (85)</td>
<td></td>
<td>0.32 ± 0.15(^d)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>16 mg</td>
<td>.49 (85)</td>
<td></td>
<td>0.62 ± 0.15(^d)</td>
</tr>
<tr>
<td></td>
<td>72</td>
<td>4 mg</td>
<td>.65 (82)</td>
<td>.47 (79)</td>
<td>0.36 ± 0.16(^d)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8 mg</td>
<td>.76 (75)</td>
<td></td>
<td>0.61 ± 0.16(^d)</td>
</tr>
<tr>
<td>Olanzapine</td>
<td>HGAP(^e)</td>
<td>10 mg</td>
<td>.24 (49)</td>
<td>.08 (49)</td>
<td>0.45 ± 0.20(^d)</td>
</tr>
<tr>
<td></td>
<td>HGAD(^f)</td>
<td>(11.6 mg)</td>
<td>.44 (62)</td>
<td>.32 (62)</td>
<td>0.23 ± 0.18(^d)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(16.4 mg)</td>
<td>.49 (65)</td>
<td></td>
<td>0.35 ± 0.18(^d)</td>
</tr>
<tr>
<td>Quetiapine</td>
<td>0004</td>
<td>250 mg</td>
<td>1.00 (6)</td>
<td>.50 (4)</td>
<td>1.51 ± 0.61(^f)</td>
</tr>
<tr>
<td></td>
<td>0006(^h)</td>
<td>(307 mg)</td>
<td>.83 (53)</td>
<td>.58 (53)</td>
<td>0.56 ± 0.19(^d)</td>
</tr>
<tr>
<td></td>
<td>0008(^i)</td>
<td>(360 mg)</td>
<td>.53 (92)</td>
<td>.37 (92)</td>
<td>0.32 ± 0.15(^d)</td>
</tr>
<tr>
<td></td>
<td>0013(^i)</td>
<td>300 mg</td>
<td>.51 (51)</td>
<td></td>
<td>0.32 ± 0.20(^d)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>600 mg</td>
<td>.45 (51)</td>
<td>.35 (51)</td>
<td>0.20 ± 0.20(^d)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>750 mg</td>
<td>.49 (53)</td>
<td></td>
<td>0.28 ± 0.20(^d)</td>
</tr>
<tr>
<td>Total</td>
<td>9 trials</td>
<td>15 arms</td>
<td>947 patients</td>
<td>529 patients</td>
<td>0.46 ± 0.06(^d)</td>
</tr>
</tbody>
</table>

\(^a\)Notes on specific studies as per Table 1.
\(^b\)Therapeutic doses are defined as risperidone ≥4 mg/day, olanzapine ≥10 mg/day, or quetiapine ≥250 mg/day. Average doses in flexible dose trials shown in parentheses. Fixed doses without parentheses.
\(^c\)Authors’ a priori definitions of response rate.
\(^d\)\(p < .05\).
\(^e\)Intent-to-treat response rates calculated from published data and compared using present investigators’ \(\chi^2\).
\(^f\)\(0.05 \leq p \leq .10\), two tailed.
\(^i\)P values were not reported by original investigators. Published data subjected to \(\chi^2\) by present investigators.

Correlation Analyses

Correlations across studies between reported placebo response rates and study design features employed weighted least squares regression as implemented in SPSS 10.0 (SPSS, Chicago), weighting response rates by the inverse of their standard errors. Response definition in this analysis employed an adjustment of the reported definitions to a common scale as shown in Table 1.

Meta-Analytic Methods

The meta-analysis features effect sizes for 15 treatment arms from nine studies. We used a two-stage procedure to account for dependencies among the effect size estimates for multiple treatment arms within studies 0204, 72, HGAD, and 0013. In the first stage we calculated a combined effect size across active dose arms for each of those studies (Gleser and Olkin 1994). This yielded nine independent study effect sizes, which were then pooled using a two-level random effects model (Bryk and Raudenbush 1992). SAS/IML (SAS, Cary, NC) software was used to calculate the combined effect sizes, and MLwiN software (Multilevel Models Project, Institute of Education, University of London, London, UK) was used to fit the random effects models. For the categorical response rates of Table 2, Cohen’s \(h\) was used to measure effect size, whereas Cohen’s \(d\) was used for the continuous response measures of Table 3. Cohen’s \(h\) and \(d\) have a comparable effect size metric in that equal values for \(h\) and \(d\) require approximately equal sample sizes for given levels of statistical power and significance (Cohen 1988).

For each response measure (categorical, continuous), random-effects models were used to estimate the “grand mean” effect size and its SE (shown in Tables 2 and 3), and to separate the variance of the nine study effect sizes into two components: the “known” sampling variances for each study and the between-study variance for the group of studies. The latter variance component reflects the consistency of the nine study effect sizes (Bryk and Raudenbush 1992). In addition, the “\(H\) statistic” (Hedges 1982; Rosenthal and Rubin 1982) was used to test a hypothesis of the homogeneity of the effect sizes.

Results

Nine placebo-controlled trials of these medications were identified: three for risperidone, two for olanzapine, and four for quetiapine. All are identified in the FDA FOI material, and all but two (Borison et al 1992; Potkin 1997) have thus far been published in full (Arvanitis and Miller 1997; Beasley et al 1996a, 1996b; Borison et al 1992, 1996; Chouinard et al 1993; Fabre et al 1995; Marder and Meibach 1994; Potkin 1997; Small et al 1997). Tables 1 and 2 summarize the designs of the nine studies. The studies employed seven different a priori response definitions. The end point change score on the BPRS was the a priori principal continuous outcome measure in 7/9 studies identified. For the two larger risperidone studies, the BPRS was considered a secondary measure and was derived from the Positive and Negative Symptom Scale (PANSS). For studies reporting multiple dose arms, each “therapeutic” dose arm is shown (Table 2), with “thera-
Table 3. Meta-Analysis of Newer Antipsychotics versus Placebo on End Point Change in Brief Psychiatric Rating Scale (BPRS)

<table>
<thead>
<tr>
<th>Newer antipsychotic</th>
<th>Studya</th>
<th>Therapeutic dose arm(s)</th>
<th>Therapeutic dose BPRS changeb mean ± SD (n)</th>
<th>Placebo BPRS changeb mean ± SD (n)</th>
<th>Effect size</th>
<th>Cohen’s d ± SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Risperidone</td>
<td>0201</td>
<td>(7.8 mg)</td>
<td>11.6 ± xx (51)</td>
<td>0.6 ± xx (53)</td>
<td>&gt;0.65</td>
<td>0.20d</td>
</tr>
<tr>
<td></td>
<td>0204</td>
<td>6 mg</td>
<td>11.3 ± 14.1 (85)</td>
<td></td>
<td>1.03</td>
<td>0.16d</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10 mg</td>
<td>5.7 ± 13.5 (85)</td>
<td>-2.2 ± 12.2 (86)</td>
<td>0.62</td>
<td>0.16d</td>
</tr>
<tr>
<td></td>
<td></td>
<td>16 mg</td>
<td>8.5 ± 15.2 (85)</td>
<td></td>
<td>0.78</td>
<td>0.16d</td>
</tr>
<tr>
<td></td>
<td></td>
<td>72</td>
<td>11.7 ± 11.5 (82)</td>
<td>8.4 ± 13.7 (79)</td>
<td>0.26</td>
<td>0.16d</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8 mg</td>
<td>12.8 ± 11.3 (75)</td>
<td></td>
<td>0.36</td>
<td>0.16d</td>
</tr>
<tr>
<td>Olanzapine</td>
<td>HGAP</td>
<td>10 mg</td>
<td>7.7 ± 12.5 (49)</td>
<td>0.2 ± 12.3 (49)</td>
<td>0.61</td>
<td>0.21d</td>
</tr>
<tr>
<td></td>
<td>HGAD</td>
<td>(11.6 mg)</td>
<td>12.6 ± 15.9 (62)</td>
<td>3.1 ± 17.5 (62)</td>
<td>0.57</td>
<td>0.18d</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(16.4 mg)</td>
<td>15.2 ± 16.1 (65)</td>
<td></td>
<td>0.73</td>
<td>0.18d</td>
</tr>
<tr>
<td>Quetiapine</td>
<td>0004</td>
<td>250 mg</td>
<td>20.9 ± 8.5 (8)</td>
<td>4.7 ± 13.2 (4)</td>
<td>1.75</td>
<td>0.73d</td>
</tr>
<tr>
<td></td>
<td>0006</td>
<td>(307 mg)</td>
<td>8.1 ± 17.4 (53)</td>
<td>2.1 ± 17.4 (53)</td>
<td>0.35</td>
<td>0.20d</td>
</tr>
<tr>
<td></td>
<td>0008</td>
<td>(360 mg)</td>
<td>8.7 ± 15.7 (92)</td>
<td>4.2 ± 15.5 (92)</td>
<td>0.29</td>
<td>0.15d</td>
</tr>
<tr>
<td></td>
<td>0013</td>
<td>300 mg</td>
<td>8.6 ± 14.7 (51)</td>
<td></td>
<td>0.71</td>
<td>0.20d</td>
</tr>
<tr>
<td></td>
<td></td>
<td>600 mg</td>
<td>7.7 ± 14.8 (51)</td>
<td>-1.7 ± 14.7 (51)</td>
<td>0.64</td>
<td>0.20d</td>
</tr>
<tr>
<td></td>
<td></td>
<td>750 mg</td>
<td>6.3 ± 14.7 (53)</td>
<td></td>
<td>0.55</td>
<td>0.20d</td>
</tr>
<tr>
<td>Total</td>
<td>9 trials</td>
<td>15 arms</td>
<td>947 patients</td>
<td>529 patients</td>
<td>&gt;0.53</td>
<td>0.07d</td>
</tr>
</tbody>
</table>

See text for meta-analytic methods. xx, not reported.

a,bPer Table 2.

*Authors’ definitions of end point change on BPRS. Minimum effect size calculated from reported inexact p value (see text).

dP < .05.

*P ≤ .10, two tailed.

*References to the primary study for each study individually. Individual study arm effect sizes ranged from 0.20 to 1.51 across arm. For the a priori defined response rate measure, these effect sizes were statistically significant for 10/15 arms (67%) using a two-tailed test and for 11/15 arms (73%) if a one-tailed test were employed. Table 3 shows the same information for the BPRS end-point change measure. For the BPRS end-point change measure, these effect sizes were statistically significant for 13/15 arms (87%) using a two-tailed test and for all 15 arms if a one-tailed test were employed.

Atypical Antipsychotic versus Placebo Comparisons: Meta-Analysis Results

Individual study arm effect sizes and results of the meta analyses are shown in Tables 2 and 3 as well. For both categorical and continuous measures, a bootstrap 95% confidence interval for the between-study variance contained zero. Additionally, the H statistic distributed as χ² (8) produced values of 7.50 (categorical) and 13.34 (continuous), which correspond to p values of approximately .48 and .10, respectively. Consistency of the between-study effect sizes is endorsed by all of these results, so a single common effect size is a reasonable representation of the nine meta-analyzed effect sizes for each measure. When drug and/or dose level were entered as covariates in the random-effects models, they reduced the between-study variance by no more than 15%.

Atypical Antipsychotic versus Placebo Comparisons: Box Score Results

The active drug versus placebo comparisons were more consistent across study than the placebo response rates. For the a priori response rate measure, Table 2 shows the individual study arm active drug versus placebo effect sizes and whether significance was reported versus placebos for each study individually. Individual study arm effect sizes ranged from 0.20 to 1.51 across arm. For the a priori defined response rate measure, these effect sizes were statistically significant for 10/15 arms (67%) using a two-tailed test and for 11/15 arms (73%) if a one-tailed test were employed. Table 3 shows the same information for the BPRS end-point change measure. For the BPRS end-point change measure, these effect sizes were statistically significant for 13/15 arms (87%) using a two-tailed test and for all 15 arms if a one-tailed test were employed.
The pooled effect size of the studies describe the estimate that the nine study samples make jointly about the outcome differences between patients who were treated with atypical antipsychotic compared to those who were treated with placebo, among the population that the studies sample. The SE describes the precision in the pooled effect size. Precision in these estimates may also be represented by confidence intervals. A 90% confidence interval based on a given sample has a 90% probability of containing the true population parameter value. There is a 10% chance that the interval lies completely above or completely below the true value, and a 5% chance that it lies completely above, under normal sampling distribution theory. Thus, the probability is 95% that the lower confidence limit for the pooled effect size for the nine studies is a lower bound for the true effect size.

For the categorical response rates, the pooled effect size of the nine studies is 0.46 with an SE of 0.06 (Table 2), and the 95% confidence lower bound of the effect size estimate is 0.36. For the continuous outcome measure, the pooled effect size is >0.53 with an SE of 0.07 (Table 3), and the 95% confidence lower bound of the effect size estimate is >0.41. The effect size estimates for the continuous outcome measure are inexact, owing to the inexact effect size denominator for study 0201.

The probability that the next atypical antipsychotic and placebo samples drawn from this population will differ significantly, when the true effect size is in fact a value greater than zero, is expressed by statistical power. Calculation of statistical power requires making an estimate of the true population effect size. To make a conservative estimate of statistical power, we conservatively set the effect size at the lower 95% confidence lower bound of our meta-analyzed estimate of the true population effect size.

With $a$ set at .05, one tailed, because the research is concerned only with the positive effect of treatment, statistical power to detect a highly conservative estimate of the effect size for $h = 0.36$ (the 95% confidence lower bound for the true effect size) for random samples of 80 per group drawn from a population of subjects similar to that of the nine studies is .67. If the grand mean effect size $h = 0.46$ is used as the estimate, power is 0.90. For the continuous outcome measure, power with the conservatively estimated effect size $d = 0.41$ is >0.82. Thus, information from the nine studies about the true population effect size suggests that we can conservatively estimate that a subsequent study with 80 patients per group will have at least an 80% chance of finding that the atypical antipsychotic is statistically superior to a placebo on a continuous outcome measure, with $a$ set at .05, using a one-tailed test. If the grand mean effect size $d = 0.53$ is used as the estimate, power is >.95.

**Subtherapeutic Doses, As Yet Unmarketed Atypical Antipsychotics, and Conventional Antipsychotics**

Several studies employed dosage arms below the therapeutic doses as defined here (Arvanitis and Miller 1997; Beasley et al 1996a, 1996b; Chouinard et al 1993; Marder and Meibach 1994; Small et al 1997). In these studies, subtherapeutic doses were superior to placebo for the *a priori* response rate measure in 17 arms and for the BPRS end point change in 27 arms.

As yet unmarketed atypical antipsychotic medications would presumably not be employed currently as standard controls, and unpublished data, if any, are not currently available through FOI requests. The published data, however, show a similar pattern of results to those from the marketed drugs. For sertindole, 3/3 BPRS comparisons of doses $\geq 20$ mg/day to placebo were statistically significant (van Kammen et al 1996; Zimbroff et al 1997). For ziprasidone, 3/3 published BPRS comparisons to placebo of doses $\geq 80$ mg/day were significant (Daniel et al 1999; Keck et al 1998).

Four of the studies reported here also included a total of six haloperidol arms (Arvanitis and Miller 1997; Beasley et al 1996b; Chouinard et al 1993; Marder and Meibach 1994; Zimbroff et al 1997), and in 6/6 arms haloperidol was significantly superior to the placebo on the BPRS end point change measure.

**Discussion**

The data suggest that the categorical placebo response rate can be quite variable in clinical trials for schizophrenia. Variation in response definition accounts for a portion of the variability across studies, with more stringent response definitions being associated with lower response rates. These correlational data across studies confirm similar findings within studies that reported multiple alternative response definitions, where more stringent definitions were consistently associated with lower response rates (Arvanitis and Miller 1997; Beasley et al 1996b; Chouinard et al 1993; Marder and Meibach 1994).

The data further suggest that the ability of the categorical response rate to distinguish active drug from placebo is also somewhat variable in trials of antipsychotic medications. For the *a priori* response definition, only two thirds of therapeutic dose arms were significantly superior to placebos, and the power of equal samples of 80 to detect a difference as statistically significant was below the conventionally accepted .80, when conservatively using the 95% confidence lower bound for the meta-analyzed population effect size estimate.

On the other hand, the dichotomous response rate is probably not the best measure to use in evaluating the consistency of the historical record of the new medications.
vis-à-vis placebos. Dichotomization of a continuous measure like the BPRS or PANSS is well known to cause considerable loss of statistical power (Cohen 1983; Kraemer 1991). The present analysis demonstrates this phenomenon in the higher effect sizes (Tables 2 and 3) and consequently higher power for the BPRS change measure than for response rate. Analyses of the BPRS as a continuous measure reveal considerable consistency in the ability of trials of newer atypical antipsychotics to show superiority of therapeutic doses to placebos, and the power of equal samples of 80 to detect a difference as statistically significant was above the conventionally accepted .80, even when conservatively using the 95% confidence lower bound for the meta-analyzed population effect size estimate.

A sample size of 80 per group is somewhat larger than the average for the nine studies reviewed, which have an average placebo sample size of 66 per group, even when the small quetiapine study 0004 is excluded. Presumably the design of the smaller studies incorporated an effect size estimate that was less conservative than the 95% confidence lower bound used here.

Modeling study as a random effect in the meta-analysis considers that the nine studies analyzed represent a random sample of studies that could potentially sample the true population. Using a random effects meta-analytic model is conservative in that random effects models assume two levels of variation, thus tending to have larger SE estimates and wider confidence intervals than their fixed effects counterparts.

The present data suggest substantial confidence that a therapeutic dose of an atypical antipsychotic would be statistically superior to a placebo in a future randomized trial, when reporting a continuous measure as the principal outcome in a study of reasonable sample size. This substantial confidence in turn suggests that historical inconsistency in the superiority of atypical antipsychotics over placebos need not be a major obstacle to considering the employment of atypical antipsychotics as active “gold standard” comparators in future equivalence trials of novel antipsychotics.

As mentioned in the introduction, however, a variety of other difficulties with the equivalence design will also need to be addressed if the equivalence design is to play a role in evaluating efficacy of novel antipsychotic compounds. One of these issues, that without placebo an active control may be biased toward greater apparent effectiveness, will be the topic of a future paper from our group. Yet another issue requiring resolution is whether and how historical data from placebo-controlled studies such as those described here would actually be used in the statistical analysis of any future equivalence trials.

Supported by USPHS Grants Nos. MH 54446 and MH 57292.

References

Carpenter WT Jr, Schooller NR, Kane JM (1997): The rationale and ethics of medication-free research in schizophrenia. Arch Gen Psychiatry 54:401–407.