Sleep improvement for restless legs syndrome patients. Part IV: meta-analysis comparison of effect sizes of vibratory stimulation sham pads and placebo pills

Fred Burbank
Salt Creek International Women's Health Foundation, San Clemente, CA, USA

Purpose: To determine whether sham pads used as controls in randomized clinical trials of vibratory stimulation to treat patients with sleep loss associated with restless legs syndrome perform differently than placebo pills used in comparable restless legs syndrome drug trials.

Patients and methods: Sham pad effect sizes from 66 control patients in two randomized clinical trials of vibratory stimulation were compared with placebo responses from 1,024 control patients in 12 randomized clinical drug trials reporting subjective sleep measurement scales. Control patient responses were measured as the standardized difference in means corrected for correlation between beginning and ending scores and for small sample sizes.

Results: For parallel randomized clinical trials, sham effects in vibratory stimulation trials were not significantly different from placebo effects in drug trials (0.37 and 0.31, respectively, $Q_{	ext{between subgroups}}=0.25, P_{Q} \geq 0.62$). Placebo effect sizes were significantly smaller in crossover drug trials than sham effect sizes in parallel vibratory stimulation trials (0.07 versus 0.37, respectively, $Q_{	ext{between subgroups}}=4.59, P_{Q} \leq 0.03$) and placebo effect sizes in parallel drug trials (0.07 versus 0.31, respectively, $Q_{	ext{between subgroups}}=5.50, P_{Q} \leq 0.02$).

Conclusion: For subjective sleep loss assessments in parallel trials, sham pads in vibratory stimulation trials performed similarly to placebo pills in drug trials. Trial design (parallel versus crossover) had a large influence on control effect sizes. Placebo pills in crossover drug trials had significantly smaller effect sizes than sham pads in parallel vibratory stimulation trials or placebo pills in parallel drug trials.

Keywords: sham effect, placebo effect, trial design, crossover study, parallel study, counterstimulation

Introduction

Background

Randomized clinical trials (RCTs) are nearly universally focused on the difference between an active treatment and an inactive or control treatment. If the active treatment effect size is large and the inactive effect size is small, the trial is a success. However, if the active treatment effect and the inactive treatment effect are both large, the trial is a failure. Clearly, the magnitude of the inactive or control treatment effect is of great importance in any RCT. In drug trials, control patients are given an inactive or a “placebo” pill that looks like the active pill but is pharmacologically inert. In physical medicine studies, control patients are exposed to a device that looks like, or is a “sham” of the physical treatment, but does not provide active treatment.
It has been suggested by Deyo et al that for drug trials, blinding of investigators and patients is generally successful.\(^1\) On the other hand, the same authors state, “Finding credible ‘placebo’ alternatives to physical therapy … may be difficult or impossible.” Kaptchuk et al examined the medical literature to determine whether or not physical medicine treatment shams have a greater therapeutic effect than drug therapy placebos, but the results were inconclusive.\(^2\) In a separate publication directly comparing the effect of a physical medicine treatment sham to a drug therapy placebo, Kaptchuk et al concluded that the sham device had a greater effect than the placebo pills on self-reported symptoms.\(^3\)

Trial designs themselves may also exert a significant influence on control effects. Deyo et al have speculated that crossover (CO) trials that expose all patients to both active and control treatments will be biased since the patients can compare their experiences with the two treatments, which may enable them to distinguish active from control treatments.\(^1\)

**Rationale**

We previously described the therapeutic effectiveness of vibratory stimulation (VS) therapy (the difference between treatment and control groups) for sleep problems associated with restless legs syndrome (RLS) and have compared that effectiveness to RLS drug therapy.\(^4,5\) We found that the magnitude of sleep improvement with a vibrating pad (treatment) was greater than with a nonvibrating (sham) pad and was comparable to sleep improvement with US Food and Drug Administration-approved drugs in patients with moderately severe primary RLS. Pad assignment (treatment pad versus sham pad) and pad assignment belief (patient belief that a treatment or sham pad was assigned) both influenced improvement in Medical Outcomes Study Sleep Problems Index II (MOS-II)\(^6-8\) sleep scores; however, pad assignment belief was more influential.\(^9\) Others have similarly reported the influence of patient belief on RCT outcomes.\(^10\)

Thus, we now examine inactive (control) effect sizes, comparing sham effect sizes in VS trials to placebo effect sizes in RLS drug trials. In addition, we will examine the influence of trial design (parallel versus CO designs) on control effect sizes and discuss the ramifications of these findings for RLS sleep studies.

**Objective**

This meta-analysis asks the question: Do sham pads used as controls in VS trials of patients with RLS sleep problems perform differently than placebos used in comparable parallel and CO RLS drug trials?

**Methods**

**RCT screening**

To compare VS sham pad effects to the published drug placebo effects reported by Fulda and Wetter, we reexamined data from control patients in two previously reported VS trials and compared them to the individual trials identified by Fulda and Wetter\(^4,5,11\).

**Measurement of placebo and sham effect sizes**

To measure placebo effect sizes of inert pills used in RLS drug trials, Fulda and Wetter calculated the magnitude of standardized mean change in sleep quality scores between baseline and endpoint separately for 1,024 control subjects across five subjective sleep instruments in 12 RLS drug trials.\(^11\) Corrections to effect sizes were made for correlation between baseline and endpoint scores and for small sample size bias.\(^12,13\) From their calculations, they arrived at placebo effect sizes on sleep scales with 95% confidence intervals (CIs). To compare subjective sleep scores for sham pad effect sizes with placebo effect sizes for inert pills, we followed the computational methods described by Fulda and Wetter\(^11\) and applied those methods to the sleep scores of 66 control patients in the VS trials.

**Statistical analysis**

**Heterogeneity testing**

Heterogeneity in treatment effect was evaluated with the \(F\) statistic (Comprehensive Meta-Analysis V2 Software, Biostat, Inc., Englewood, NJ, USA).\(^14\) \(F\) values ranged from 0% to 100%, with ≤25%, 50%, and ≥75% corresponding to low, medium, and high heterogeneity, respectively.\(^15\) To compensate for insensitivity of the \(F\) statistic for small sample sizes, when \(P__{\chi^2}<0.10\), the null hypothesis of homogeneity was rejected and studies were considered heterogeneous. For all other statistical tests, significance cutoff was at \(P\leq0.05\).

**Meta-analysis models**

Outcome measures were directly compared by random-effects statistical models. Subgroups were indirectly compared using the \(Q\) statistic.\(^14\)

**Measurement of sleep improvement**

For the VS trials, sleep problems were measured with the MOS-II sleep problems index.\(^7,8\) Differences in MOS-II
scores from baseline to endpoint (change scores) were calculated for sham groups and converted to standardized mean differences using the baseline standard deviations. Change scores were also corrected for baseline and endpoint correlation and for small sample size bias. The MOS sleep inventory is a patient-reported, 12-question, paper-and-pencil test that has been shown to be reliable and valid for measuring sleep problems in patients with RLS. The MOS-II scale contains 9 of the 12 inventory questions, represents all of the qualitative sleep concepts in the inventory, and reflects the inventory’s most exhaustive measure of sleep problems. In the current analysis, which follows the Fulda and Wetter convention, improvement in a sleep score was calculated as a positive number, indicating a reduction in sleep difficulty; the greater the positive number, the greater the subjective sleep improvement.

For the 12 drug trials, five different subjective measures of sleep quality were included: 1) the two-question “sleep adequacy” items in the MOS sleep inventory, 2) the Schlaffragebogen A “sleep quality” scale, 3) the “satisfaction with sleep” item of the RLS-6 scale, 4) a visual analog “satisfaction with sleep” scale, and 5) a diary-derived sleep quality scale.

Null hypotheses tested
Hypothesis for control types (inactive sham pads versus inert pills):
Efficacy of VS sham pads compared to drug placebos (indirect subgroup comparisons)
\[ H_{01} : S\Delta MC_{\text{in sleep quality scores}} = S\Delta MC_{\text{in sleep quality scores}} \]
for sham pads
for drug placebos.
Hypothesis for trial designs (parallel versus CO):
Efficacy of parallel RCT compared to CO trials (indirect subgroup comparison)
\[ H_{02} : S\Delta MC_{\text{in sleep quality scores}} = S\Delta MC_{\text{in sleep quality scores}} \]
for parallel RCTs
for CO trials
where \( S\Delta MC \) is the standardized difference in mean change between initial and final sleep quality scores, corrected for initial and final score correlation and for small sample sizes.

Results
Trials selected
Details of the placebo analysis of 12 RLS drug trials and the two VS trials have been previously published. Controls for the drug trials were pharmacologically inert pills identical in appearance to study drugs. Controls for the VS trials were non-vibrating pads that were identical in appearance to vibrating pads, but which did not produce vibration. In SMI-001 the sham pads produced patient-controlled sound; in SMI-002, patient-controlled light. The VS trials demonstrated low heterogeneity (\( F=0.0\% \)); the 12 drug trials, moderate heterogeneity (\( F=70.7\% , \ P_0<0.0001 \)). Drug trial heterogeneity was not a function of trial date, parallel versus CO trial design, trial size, drug studied, or subjective sleep scale used and could not be explained by our analysis of the published data.

Sensitivity analyses
Trials with many different subjective sleep indices, all but one of which have not been validated in RLS populations, were included in the analysis. To determine whether non-validated sleep scales exerted an influence, control effect sizes for the eight trials that used validated MOS subscales were compared with the remaining trials that did not use a validated sleep scale. No significant difference in control effect size was seen between the trials that used MOS subscales and those that did not (0.270 versus 0.304, respectively, \( P_0 \geq 0.69 \)). Similarly, trials varied considerably in the size of patient enrollment. However, meta-regression of control effect size as a function of control patient enrollment demonstrated no significant relationship (slope \( =0.0004, \ P \geq 0.24 \)).

Outcomes
Figure 1 is a forest plot of effect sizes, with CIs for individual trials and for three trial subgroups: CO drug trials, parallel drug trials, and parallel VS trials. Improvement in subjective sleep quality scores is shown as positive values. Control effect sizes were significantly greater than zero for the parallel drug trials (\( P \leq 0.0001 \), Table 1) and for parallel VS trials (\( P \leq 0.0001 \), Table 1). Although control effect sizes were slightly larger for shams than for inert pills, 0.365 and 0.308, respectively, the difference between these two parallel trial subgroups was not significant (\( P_0 \geq 0.62 \), Table 2). Therefore, hypothesis \( H_{01} \) was accepted for parallel VS trials compared to parallel drug trials.

In contrast to the parallel studies, control effect sizes were not significantly different than zero for the CO drug trials (0.073, \( P \geq 0.41 \), Table 1). In these trials, placebo had no significant therapeutic effect. When sham effect sizes in the parallel VS trials were compared with placebo effect sizes in CO drug trials, placebo effects in CO drug trials were significantly smaller (0.365 versus 0.073, respectively, \( P_0 \leq 0.03 \), Tables 1 and 2). Therefore, hypothesis \( H_{01} \) was
rejected for parallel VS trials compared with CO drug trials. Placebos in CO drug trials also had significantly smaller effect sizes than placebos in parallel drug trials (0.073 versus 0.308, respectively, \( P_0 \leq 0.02 \), Tables 1 and 2). Therefore, hypothesis \( H_0 \) was rejected for parallel drug trials compared to CO drug trials.

**Discussion**

For parallel trial designs, control treatment effect sizes were quite similar for placebo pills and sham pads (0.308 and 0.365, respectively, \( P_0 \geq 0.62 \)). Indirect comparisons of RLS drug and VS trials are, therefore, justified, so long as the compared trials are of parallel design. For example, the indirect comparison of VS and drug studies made by Burbank et al would be valid because the compared trials were both parallel designs.

By contrast, placebo groups in CO drug trials had little or no therapeutic effect on subjective sleep measures (effect size =0.073, −0.1000 to 0.246, 95% CI). The lack of significant therapeutic effect in the control arms of these studies raises the suspicion that blinding was unsuccessful. If so, it is impossible to make indirect comparisons of these trials with other studies. Moreover, the primary, direct comparison of their respective treatment arms and control arms comes into question.

It appears that patients in the CO drug trials identified which treatment was the control treatment and which the active one. In the beginning of these trials, half of the patients were randomized to a drug and half were randomized to a placebo. At the half-way point, a week-long “washout” period was inserted, during which no treatment was given.

**Table 1** Subgroup effect sizes for CO drug trials, parallel drug trials, and parallel VS trials

<table>
<thead>
<tr>
<th>Trial subgroup</th>
<th>Number of trials</th>
<th>Number of patients</th>
<th>Effect size</th>
<th>( P )-value</th>
<th>Significant change from baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>CO drug trials</td>
<td>3</td>
<td>66</td>
<td>0.073</td>
<td>0.409</td>
<td>No</td>
</tr>
<tr>
<td>Parallel drug trials</td>
<td>9</td>
<td>958</td>
<td>0.308</td>
<td>( \leq 0.0001 )</td>
<td>Yes</td>
</tr>
<tr>
<td>Parallel VS trials</td>
<td>2</td>
<td>66</td>
<td>0.365</td>
<td>( \leq 0.0001 )</td>
<td>Yes</td>
</tr>
</tbody>
</table>

**Table 2** Pairwise subgroup effect size comparisons between trial subgroups

<table>
<thead>
<tr>
<th>Subgroup comparison</th>
<th>( Q )-between subgroups</th>
<th>( P_0 )-value</th>
<th>Significant difference between subgroups</th>
</tr>
</thead>
<tbody>
<tr>
<td>CO drug trials versus parallel drug trials</td>
<td>5.496</td>
<td>0.019</td>
<td>Yes</td>
</tr>
<tr>
<td>CO drug trials versus parallel VS trials</td>
<td>4.588</td>
<td>0.032</td>
<td>Yes</td>
</tr>
<tr>
<td>Parallel drug trials versus parallel VS trials</td>
<td>0.252</td>
<td>0.616</td>
<td>No</td>
</tr>
</tbody>
</table>

**Abbreviations:** CO, crossover; VS, vibratory stimulation.
Following the washout period, patients who were initially randomized to a drug were given the placebo, and patients initially randomized to the placebo were given the active drug. Because all the drugs that were studied have soporific effects that occur within an hour or so after ingestion, patients may have been able to distinguish the active drug from the placebo following CO. For those who received the active drug in the first half of the trial, it is likely that loss of the soporific effect signaled to them that they were receiving the placebo in the second half, which biased them toward reporting non-improvement following CO. Similarly, for those who had received a placebo in the first half of the trial, receiving a pill that caused sleepiness may have suggested to them that they were receiving the active drug in the second half, which biased them to report improvement following CO.

Accurate patient beliefs about the drugs received following CO could have influenced trial results. As previously demonstrated, once RLS patients develop a belief about the type of treatment they received (active versus control), their sleep inventory scores are strongly influenced. Patients who believed they had been given a placebo reported little sleep improvement. Patients who believed they were given the active treatment reported substantial sleep improvement.

It is possible that CO designs in RLS drug trials can only maintain adequate patient blinding if the active treatment has no discernable effect (in these studies, sleepiness) or the placebo has a soporific effect that is comparable to that of the active treatment, rather than being a completely inert pill. In CO studies, bias toward treatment might also be minimized by using a different outcome measurement, such as objectively recording sleep efficiency in a sleep laboratory, rather than relying on patient-reported, subjective sleep measurement scales.

With their presentation of independent standardized effect sizes for control groups, Fulda and Wetter created a statistic that may help evaluate blinding in RLS trials. The average drug trial placebo effect size of 0.308 (0.215–0.401 95% CI) and the average VS trial sham effect size of 0.365 (0.161–0.569 95% CI) set a relatively narrow range of values against which any new study of sleep disturbance in RLS patients could be judged. If, for example, a new study had a control effect size that was zero or nearly zero, as reported in three of the Fulda and Wetter CO trials and in two of their parallel trials, one might suspect that sometime during the course of the trial, patients discerned whether they received the active or the control treatment. It would seem unreasonable to assume that trial blinding was successful if a trial had a control effect size of zero or nearly zero. Of course, additional measures should be used to evaluate blinding in any trial, such as measuring compliance with treatment and follow-up schedules or determining which study arm each patient guessed they had been assigned to using questionnaires. However, such additional measures aside, simply examining standardized effect size for control patients may provide useful clues about blinding adequacy.

In addition to the well-known general limitations of meta-analyses, limitations specific to the current meta-analysis exist. The primary limitation is the fact that we included only the drug trials chosen by Fulda and Wetter. Those trials were moderately heterogeneous, which could not be explained by Fulda and Wetter nor by us. Heterogeneity argues against trials being integrated though meta-analysis. However, our decision to use the same set of previously selected trials allowed comparison to a known, published standard. Another limitation is the examination of only one outcome variable: sleep problems. It may well be that other measurements of discomfort in RLS patients would not follow the same patterns observed in sleep inventories. In addition to showing that sleep loss measures for sham patients were smaller in CO trials than in parallel trials, Fulda and Wetter also demonstrated that effect sizes for RLS severity measures were smaller in CO trials than in parallel trials. This observation suggests that our results may not be limited to sleep problem measures alone.

Conclusion

Sham pads in parallel VS trials performed similarly to placebo pills in parallel drug trials for subjective sleep loss assessments in patients with RLS. Therefore, so long as RLS trials are of parallel design, subjective sleep loss assessment in VS trials can be indirectly compared to sleep problems in drug trials through meta-analysis. Trial design (parallel versus CO) influenced control effect sizes. Placebo pills in CO drug trials had significantly smaller effect sizes than sham pads in parallel randomized VS trials and significantly smaller effect sizes than placebo pills in parallel drug trials. CO trial designs in the study of subjective measures of sleep problems in RLS patients may be biased toward showing a treatment effect because patients may be able to discern treatment from placebo, causing control arms in CO trials to have little or no therapeutic effect.

Disclosure

Financial support for the study was provided by Sensory Medical, Inc, San Clemente, CA, USA. The author is the
Chief Executive Officer of Sensory Medical, Inc, and a minority shareholder. The author reports no other conflicts of interest.

References