ORIGINAL RESEARCH ARTICLE

Do Unblinded Assessors Bias Muscle Strength Outcomes in Randomized Controlled Trials of Progressive Resistance Strength Training in Older Adults?

ABSTRACT


Objective: Knowledge of treatment assignment and failing to analyze results by randomized treatment groups—an intention-to-treat analysis—may cause bias in the treatment effect estimate in randomized controlled trials. This study was undertaken to determine the difference in lower limb muscle strength measured by blinded vs. by unblinded outcome assessors in 73 progressive resistance strength training trials conducted in older adults.


Results: Meta-regression analyses showed that trials that used blinded assessors \( n = 18 \) tend to report smaller effect sizes than do those that used unblinded assessors \( n = 55 \), with a difference of \(-0.80\) (95% confidence interval, \(-1.35\) to \(-0.25\)). This result still holds even after adjusting for the use of an intention-to-treat analysis, with an adjusted difference of \(-0.65\) (95% confidence interval, \(-1.26\) to \(-0.04\)). The reported effects were exaggerated in trials that used unblinded assessors.

Conclusions: This study suggests that assessor blinding is important and is a safeguard to the internal validity of exercise trials in older adults.

Key Words: Muscle Strength, Bias (Epidemiology), Single-Blind Method, Randomized Controlled Trial
A large body of studies has shown that progressive resistance strength training can restore age-related loss in muscle strength and improve physical functioning. The effect of such training is promising even in older adults with chronic conditions such as osteoarthritis that may create long-term disability. Many of these studies have applied a randomized controlled trial (RCT) design, which is assumed to yield the highest level of evidence. An RCT design is considered to be a robust research design because it reduces threats to the internal validity of the study through the randomization procedure, in which the treatment assignment for each participant is determined by chance. The purpose of the randomization procedure is to eliminate any known or unknown potential confounding factors so that the control group and the experimental group are comparable before intervention. However, this sound design may still yield flawed results if detection bias from participants or outcome assessors is not controlled.

Despite randomization, bias can be introduced into the results of an RCT through inadequate blinding of participants and outcome assessors. Blinding, which is used to prevent detection bias during the outcome assessment, is different from concealment of treatment allocation, which is used to avoid selection bias during the randomization process. Blinding is done because both participants and assessors could be influenced by the knowledge about treatment assignment and induce a systematic detection bias. Spurious treatment effects have been found in unblinded RCTs that measured either continuous outcomes or dichotomous outcomes. Overestimation of treatment effect resulting from unblinded assessors has been estimated to range from 17% to 40%.

While treatment effects may be overestimated by using unblinded assessors, they may be further overestimated by failure to use an intention-to-treat (ITT) analysis. The ITT analysis is based on randomly determined treatment assignments instead of treatment receipt or adherence. If a participant was assigned to the intervention group but deviated from treatment protocol (i.e., poor adherence), this participant’s outcome data are still analyzed, with all others randomized to the intervention under the principle of ITT analysis. The purpose of conducting an ITT analysis is to safeguard the random allocation, which is presumed to balance the known and unknown characteristics of participants that may affect the treatment outcome, from events that occur after the randomization. Estimates from an ITT analysis are also considered to be more approximate to treatment effects in a clinical setting because not all patients will completely comply with a treatment regimen. A high-quality RCT that applies the principle of blinding to evaluate outcomes can also use an ITT analysis to analyze the outcome data. Both blinding and ITT analysis approaches help prevent the overestimation of an intervention effect. Previous studies that assessed the bias effect often focus on the single factor of blinding and overlook the possibility that the effect of blinding may be confounded with the effect of an ITT analysis. The observed treatment effect, in fact, might be a result of combining the blinding and ITT approaches. Untangling the individual influence of blinding and ITT is necessary especially when conducting a meta-analysis to pool effects from multiple trials, not all of which have blinded outcome assessors or ITT analyses. This further analysis can help explain why some trials have larger effects than others do. This further inquiry during a meta-analysis is important especially when conflicting trial results are found.

Inconsistent research methodology and statistical analysis used in RCTs may contribute to heterogeneity in a meta-analysis. In addition, a single overall estimate of treatment effect may be too simple because the conduct of a meta-analysis is retrospective and susceptible to several sources of bias. Investigating the potential sources of heterogeneity when conducting a meta-analysis is highly recommended. Identification of the source of heterogeneity may explain the difference in effect sizes observed among trials. This can help clinicians interpret meta-analysis results and make decisions applying research results into practice when adopting an evidence-based practice. Although there is mounting evidence showing that muscle strength training, especially progressive resistance strength training, can improve muscle strength in older adults, unblinded trials may overstate the effect.

Unlike pharmacologic trials, in which blinding can be done by using drugs and placebos that have an identical appearance, blinding of participants in muscle strength training trials is problematic because the participants need to take part in the intervention. Alternatively, it is possible to have blinded assessors in these trials to objectively assess outcomes. The purpose of this study was to determine the differential effects of progressive resistance strength training on lower limb muscle strength in older adults between RCTs that used blinded outcome assessors and those that did not. As a further step, we will account for the influence of blinding and ITT approaches on treatment effects.
of ITT analysis while estimating the effect of blinding through a meta-regression analysis.

METHODS

The trials analyzed in the current study have been recently published in a systematic review by the Bone, Joint, and Muscle Trauma Group of the Cochrane Collaboration, in which only RCTs were included.\(^1\) Trials in that systematic review were screened by two authors (C.-J. Liu and N.K. Latham) individually with the following inclusion criteria: (1) the research design is an RCT; (2) the mean age of the trial participants is 60 yrs or older; and (3) progressive resistance strength training was the primary intervention exercise. An additional criterion was added for the purpose of the current study—the trial outcome measure must include lower limb muscle strength. A trial was considered to have blinded outcome assessors or an ITT analysis only if related information could be clearly identified in the method and result sections of the published text. The information about assessor blinding and ITT analysis is usually described with words such as *blind* or *mask* and *intention-to-treat* or *intend-to-treat*. If there was ambiguity in the trial report, the trial was considered to have unblinded outcome assessors or non-ITT analysis. The same two authors independently extracted information about blinding and ITT analysis into a standardized paper form. If disagreement occurred, the two authors would meet to reach consensus through discussion.

We calculated the treatment effect (effect size) of each trial as the standardized mean difference (SMD) between study arms, because muscle strength was measured as a continuous outcome. The SMD is the difference in means between the intervention group and the control group divided by the pooled standard deviation of lower limb muscle strength in the intervention and control groups. Because the SMD gives the difference between groups in standard deviation units, this effect size does not depend on the measurement scale of muscle strength, such as kilograms or pounds from the one-repetition maximum measure or newton meters from the kinematics measure. To see whether the SMDs were more dispersed than would be expected (heterogeneity), the Cochran \(Q\) test was used to test for excess variability. A random-effects meta-analysis model was used to calculate the summary effect size. \(^2\) Egger test was used to evaluate the study results for publication bias.\(^22\)

To estimate the influence of blinding on the summary effect while accounting for the impact of an ITT analysis, we tried two approaches. In the first approach, we partitioned the trials into four subgroups: blinded trials with an ITT analysis, blinded trials without an ITT analysis, unblinded trials with an ITT analysis, and unblinded trials without an ITT analysis. The summary effect sizes and 95% confidence intervals (CIs) for these four subgroups were separately obtained using random-effects meta-analysis and compared. In the second approach, to include all of the trials within a single analysis, we used a random-effects meta-regression model with both the blinding of the outcome assessors and the use of ITT analysis as covariates.\(^19\,\,\,23\) The meta-regression analysis was performed with the SAS 9.1 statistical software using the MIXED procedure.\(^24\)

RESULTS

Seventy-three trials (3059 cases) were included in this study. The sample mean age was 72.4 yrs. Twenty-five trials recruited older adults without preexisting health conditions; 20 recruited older adults with a health condition such as osteoarthritis, coronary heart disease, or hip replacement; 15 recruited older adults with a sedentary lifestyle; and 13 recruited frail older adults or those who had functional limitations or weak leg muscle strength. Eighteen trials reported using blinded outcome assessors, and 11 reported using an ITT analysis. Among these trials, nine reported using both.

<table>
<thead>
<tr>
<th>Subgroup (Number of Trials)</th>
<th>Effect Size (95% CI)</th>
<th>(Q) Test ((P))</th>
<th>Egger Test ((P))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blinded with an ITT ((n = 9))</td>
<td>0.12 (-0.01 to 0.25)</td>
<td>(\chi^2 = 17\ (0.03))</td>
<td>0.01 (0.62)</td>
</tr>
<tr>
<td>Blinded without an ITT ((n = 9))</td>
<td>0.54 (0.23-0.84)</td>
<td>(\chi^2 = 5.91\ (0.66))</td>
<td>0.00 (0.93)</td>
</tr>
<tr>
<td>Unblinded with an ITT ((n = 2)^a)</td>
<td>1.07 (0.50-1.64)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Unblinded without an ITT ((n = 53))</td>
<td>1.05 (0.83-1.27)</td>
<td>(\chi^2 = 191\ (&lt;0.01))</td>
<td>0.00 (0.91)</td>
</tr>
</tbody>
</table>

CI indicates confidence interval; ITT, intention-to-treat analysis.

\(^a\)\(Q\) and Egger tests are not performed because of the small number of trials.
Trial SMD ranged from −0.58 to 5.31. The Egger test showed no evidence of publication bias (P = 0.39), but the Q test showed significant heterogeneity between trials ($\chi^2 = 293.86, df = 72, P < 0.001$). The random-effects model produced a large pooled effect size (SMD, 0.84; 95% CI, 0.69–1.00), which indicates a strong effect of progressive resistance strength training on lower limb muscle strength in older adults. After separating trials into four subgroups, there was no significant effect of progressive resistance strength training in blinded trials with an ITT analysis. Blinded trials without an ITT analysis showed a moderate effect. Unblinded trials showed a large effect regardless of whether using an ITT analysis or not. The SMDs and 95% CIs for the four subgroups of trials are reported in Table 1, along with the Q and Egger test results.

When assessor blinding was the only covariate in the meta-regression model, trials that reported using blinded outcome assessors had smaller effect sizes than did trials that did not (mean difference in SMD between the blinded and unblinded trials, −0.80; 95% CI, −1.35 to −0.25). The estimated SMD was 1.22 for the blinded trials and 0.42 for the unblinded trials. Similarly, when ITT analysis was the only covariate in the model, trials that reported using ITT analysis had smaller effect sizes than did trials that did not (mean difference in SMD between trials that used an ITT analysis and trials that did not, −0.74; 95% CI, −1.41 to −0.01). The estimated SMD was 1.14 for trials that used an ITT analysis and 0.40 for trials that did not. Figures 1 and 2 are the bubble graphs that show the difference in SMD of trials by the covariate of blinding and ITT. Note that the size of each bubble is a function of the inverse variance of each trial; accordingly, larger bubbles are the trials with small variance, which were assigned larger weight in the meta-analysis and also the meta-regression analysis. When both covariates, assessor blinding and ITT, were controlled for simultaneously, the result of assessor blinding still held significance, whereas ITT analysis

\[ \text{FIGURE 1} \quad \text{Bubble graph of assessor blinding.} \]

\[ \text{FIGURE 2} \quad \text{Bubble graph of intention-to-treat analysis.} \]
was no longer significant (mean difference in SMD, 
\(-0.65; 95\% \text{ CI, } -1.26 \text{ to } -0.04\); and 
\(-0.39; 95\% \text{ CI, } -1.11 \text{ to } 0.25\), respectively). The estimated SMD 
was 0.86 for unblinded trials vs. 0.21 for blinded 
trials after adjustment for ITT analysis. This finding shows that detection bias has a significant influence on the effect of lower limb muscle strength in these trials.

**DISCUSSION**

Blinding of outcome assessors is one methodologic safeguard to ensure the internal validity of an RCT. Using blinded assessors is critical in exercise trials to yield unbiased outcomes because blinding of participants cannot be done easily in these trials. However, lack of blinded assessors in the progressive resistance strength training trials that we identified is common (75% of 73 trials). The outcome of lower limb muscle strength is exaggerated in trials that did not use a blinded outcome assessor, based on our subgroup analysis and meta-regression analysis. Although the estimate of blinding is slightly reduced after the use of an ITT analysis has been taken into account, according to our estimation, the magnitude of intervention effect is still four times more in unblinded trials than in blinded trials.

Progressive resistance strength training has demonstrated a positive effect to reverse age-related loss in muscle strength; however, trials that failed to use blinded assessors have shown larger apparent effects than trials that used blinded assessors. In our study, this discrepancy was not accounted for by the use of an ITT analysis, which might also reduce the apparent treatment effect. These findings suggest that unblinded assessors are unable to assess the outcome of muscle strength objectively. Although the intervention effect on lower limb muscle strength is positive regardless of the use of blinded assessors, ignoring the influence of unblinded assessors can obscure the true effect. Our study indicates that using unblinded outcome assessors can lead to overestimation of the intervention effect by a factor of four. This finding is consistent with the findings of previous studies evaluating other types of interventions that found that unblinded trials tend to yield spurious results.\(^8\)\(^{-13}\) If trial investigators are interested in the benefit/harm ratio of an exercise intervention, the result will not be trustworthy when the assessor is not blinded. Clinical practice guidelines developed based on such unblinded trials would be misleading.

Although muscle strength is a performance-based measure, unblinded outcome assessors can be a threat to the validity of the results. Objective outcome measures, such as death, are less affected by assessors, whereas subjective outcome measures, such as self-report measures, are.\(^7\)\(^,{13}\) Although measurements obtained from physical performance tests seem objective, detection bias from unblinded outcome assessors is possible. Guyatt et al.\(^25\) found that simple encouragement, such as “You’re doing well” or “Keep up the good work,” from the assessor could significantly increase patients’ walking distance in the 6-min walk test. Therefore, if an outcome assessor is not blinded to treatment allocation, the knowledge of treatment may tempt the assessor to prompt participants’ performance in the experimental group and produce spurious results that are not contributed by the intervention.

Although blinding of participants in exercise trial is problematic, unless there is some level of regular exercise or sham exercise in the control group, blinding of outcome assessors is easier. Lack of blinded assessors, however, is common in exercise trials.\(^26\) Our study has identified that 55 of 75 RCTs did not report the use of blinded outcome assessors. This methodologic weakness downgrades the credibility of study results. Lack of blinded outcome assessors is also common in rehabilitation studies. Johnston et al. reviewed rehabilitation research in spinal cord injuries, traumatic brain injuries, and burn injuries from 1999 to 2004 and found lack of blinded assessors to measure outcomes in most identified RCTs.\(^2\) Like exercise trials, participants in rehabilitation trials often need to actively take part in the treatment process. Blinding of participants will need some type of sham intervention to disguise the control group. We do not suggest that exercise and rehabilitation researchers should not make an effort to blind participants because it is hard to achieve. Instead, we encourage researchers to blind participants if possible. As blinding of assessors is relatively feasible, we highly recommend researchers to take this step to increase the internal validity of the study.

An RCT design does not simply produce high-quality research if the internal validity is not carefully safeguarded. As our study demonstrated, treatment effects are exaggerated in trials that used unblinded assessors. In fact, our study might have underestimated the blinding effect. We might have misclassified some blinded trials into unblinded trials if the trial authors did not report blinding-related information in the published text. It is possible that some heterogeneity found in the unblinded trials without an ITT subgroup is caused by poor reporting of the trial rather than poor research methodology, which could be a limitation to our...
study. This group could be a mix of trials that had blinded assessors but failed to report the blinding and studies without blinded assessors. The Consolidated Standards for Reporting of Trials Group, in the most recent statement, has suggested that authors should report whether blinding was applied and also describe the blinding effort (item 11a). Blinding effort may include the use of sham intervention to the control group or prohibition of discussion about interventions between participants and outcome assessors. The investigators should validate the effectiveness of blinding by asking assessors to guess actual treatment assignments at the end of the trial. Current blinded trials usually provided little information in their report on how the assessors were blinded to the treatment assignment. Only a small fraction of trials in top medical and psychiatry journals reported evidence of successful blinding. Journal reviewers and editors should encourage investigators to report such information to increase publication quality.

High-quality studies are desirable in the field of rehabilitation to provide rigorous evidence for interventions. Accurately determining the effectiveness of a rehabilitation intervention from multiple research trials is important to guide clinical practice. Sometimes, this task becomes difficult because of the heterogeneity that exists in studies addressing the same therapeutic intervention. Although different clinical populations and intervention programs may cause heterogeneity, we did not find the participants’ health status or the training intensity to be significant in our preliminary analyses (unpublished). Our study has shown that the inconsistency of blinding greatly moderates the magnitude of effect size of lower limb muscle strength outcomes in progressive resistance strength training trials in older adults. Knowledge of treatment assignment might bias assessors. To safeguard the internal validity of an RCT, especially in an exercise trial or rehabilitation trial when blinding of participants may not be easily done or when subjective outcomes are assessed, blinding of assessors is the fundamental approach. Moreover, to help readers interpret and appraise study results, trial authors are encouraged to follow the Consolidated Standards of Reporting Trials (CONSORT) statement in reporting the blinding efforts taken as part of a clinical trial.

REFERENCES


Unblinded Assessors and Bias in RCTs