

Supplementary S1: Quality Assessment Checklist and Guidelines

Study:

Criteria (Yes/No/CD/NR/NA)	Rater 1	Rater 2	Rater 3	Revised
1. Was the study described as randomized, a randomized trial, a randomized clinical trial, or an RCT?				
2. Was the method of randomization adequate (i.e., use of randomly generated assignment)?				
3. Was the treatment allocation concealed (so that assignments could not be predicted)?				
4. Were study participants and providers blinded to treatment group assignment?				
5. Were the people assessing the outcomes blinded to the participants' group assignments?				
6. Were the groups similar at baseline on important characteristics that could affect outcomes (e.g., demographics, risk factors, co-morbid conditions)?				
7. Was the overall drop-out rate from the study at endpoint 20% or				

Supplementary S1: Quality Assessment Checklist and Guidelines

Criteria (Yes/No/CD/NR/NA)	Rater 1	Rater 2	Rater 3	Revised
lower of the number allocated to treatment?				
8. Was the differential drop-out rate (between treatment groups) at endpoint 15 percentage points or lower?				
9. Was there high adherence to the intervention protocols for each treatment group?				
10. Were other interventions avoided or similar in the groups (e.g., similar background treatments)?				
11. Were outcomes assessed using valid and reliable measures, implemented consistently across all study participants?				
12. Did the authors report that the sample size was sufficiently large to be able to detect a difference in the main outcome between groups with at least 80% power?				
13. Were outcomes reported or subgroups analyzed prespecified				

Supplementary S1: Quality Assessment Checklist and Guidelines

Criteria (Yes/No/CD/NR/NA)	Rater 1	Rater 2	Rater 3	Revised
(i.e., identified before analyses were conducted)?				
14. Were all randomized participants analyzed in the group to which they were originally assigned, i.e., did they use an intention-to-treat analysis?				
Rater	Quality Rating (Good, Fair, or Poor)	Additional Comments (If POOR, please state why):		
SCP				
PM				
BP				
My Revised Rating:				

*CD, cannot determine; NA, not applicable; NR, not reported

Supplementary S1: Quality Assessment Checklist and Guidelines

Guidance for Assessing the Quality of Controlled Intervention Studies

The guidance document below by the National Heart Lung and Blood Institute (2013) is organized by question number from the tool for quality assessment of controlled intervention studies.

Question 1. Described as randomized

Was the study described as randomized? A study does not satisfy quality criteria as randomized simply because the authors call it randomized; however, it is a first step in determining if a study is randomized

Questions 2 and 3. Treatment allocation—two interrelated pieces

Adequate randomization: Randomization is adequate if it occurred according to the play of chance (e.g., computer generated sequence in more recent studies, or random number table in older studies).

Inadequate randomization: Randomization is inadequate if there is a preset plan (e.g., alternation where every other subject is assigned to treatment arm or another method of allocation is used, such as time or day of hospital admission or clinic visit, ZIP Code, phone number, etc.). In fact, this is not randomization at all—it is another method of assignment to groups. If assignment is not by the play of chance, then the answer to this question is no.

There may be some tricky scenarios that will need to be read carefully and considered for the role of chance in assignment. For example, randomization may occur at the site level, where all individuals at a particular site are assigned to receive treatment or no treatment. This scenario is used for group-randomized trials, which can be truly randomized, but often are "quasi-experimental" studies with comparison groups rather than true control groups. (Few, if any, group-randomized trials are anticipated for this evidence review.)

Allocation concealment: This means that one does not know in advance, or cannot guess accurately, to what group the next person eligible for randomization will be assigned. Methods include sequentially numbered opaque sealed envelopes, numbered or coded containers, central randomization by a coordinating center, computer-generated randomization that is not revealed ahead of time, etc.

Supplementary S1: Quality Assessment Checklist and Guidelines

Questions 4 and 5. Blinding

Blinding means that one does not know to which group–intervention or control–the participant is assigned. It is also sometimes called "masking." The reviewer assessed whether each of the following was blinded to knowledge of treatment assignment: (1) the person assessing the primary outcome(s) for the study (e.g., taking the measurements such as blood pressure, examining health records for events such as myocardial infarction, reviewing and interpreting test results such as x ray or cardiac catheterization findings); (2) the person receiving the intervention (e.g., the patient or other study participant); and (3) the person providing the intervention (e.g., the physician, nurse, pharmacist, dietitian, or behavioral interventionist).

Generally placebo-controlled medication studies are blinded to patient, provider, and outcome assessors; behavioral, lifestyle, and surgical studies are examples of studies that are frequently blinded only to the outcome assessors because blinding of the persons providing and receiving the interventions is difficult in these situations. Sometimes the individual providing the intervention is the same person performing the outcome assessment. This was noted when it occurred.

Question 6. Similarity of groups at baseline

This question relates to whether the intervention and control groups have similar baseline characteristics on average especially those characteristics that may affect the intervention or outcomes. The point of randomized trials is to create groups that are as similar as possible except for the intervention(s) being studied in order to compare the effects of the interventions between groups. When reviewers abstracted baseline characteristics, they noted when there was a significant difference between groups. Baseline characteristics for intervention groups are usually presented in a table in the article (often Table 1).

Groups can differ at baseline without raising red flags if: (1) the differences would not be expected to have any bearing on the interventions and outcomes; or (2) the differences are not statistically significant. When concerned about baseline difference in groups, reviewers recorded them in the comments section and considered them in their overall determination of the study quality.

Supplementary S1: Quality Assessment Checklist and Guidelines

Questions 7 and 8. Dropout

"Dropouts" in a clinical trial are individuals for whom there are no end point measurements, often because they dropped out of the study and were lost to followup.

Generally, an acceptable overall dropout rate is considered 20 percent or less of participants who were randomized or allocated into each group. An acceptable differential dropout rate is an absolute difference between groups of 15 percentage points at most (calculated by subtracting the dropout rate of one group minus the dropout rate of the other group). However, these are general rates. Lower overall dropout rates are expected in shorter studies, whereas higher overall dropout rates may be acceptable for studies of longer duration. For example, a 6-month study of weight loss interventions should be expected to have nearly 100 percent followup (almost no dropouts—nearly everybody gets their weight measured regardless of whether or not they actually received the intervention), whereas a 10-year study testing the effects of intensive blood pressure lowering on heart attacks may be acceptable if there is a 20-25 percent dropout rate, especially if the dropout rate between groups was similar. The panels for the NHLBI systematic reviews may set different levels of dropout caps.

Conversely, differential dropout rates are not flexible; there should be a 15 percent cap. If there is a differential dropout rate of 15 percent or higher between arms, then there is a serious potential for bias. This constitutes a fatal flaw, resulting in a poor quality rating for the study.

Question 9. Adherence

Did participants in each treatment group adhere to the protocols for assigned interventions? For example, if Group 1 was assigned to 10 mg/day of Drug A, did most of them take 10 mg/day of Drug A? Another example is a study evaluating the difference between a 30-pound weight loss and a 10-pound weight loss on specific clinical outcomes (e.g., heart attacks), but the 30-pound weight loss group did not achieve its intended weight loss target (e.g., the group only lost 14 pounds on average). A third example is whether a large percentage of participants assigned to one group "crossed over" and got the intervention provided to the other group. A final example is when one group that was assigned to receive a particular drug at a particular dose had a large percentage of participants who did not end up taking the drug or the dose as designed in the protocol.

Supplementary S1: Quality Assessment Checklist and Guidelines

Question 10. Avoid other interventions

Changes that occur in the study outcomes being assessed should be attributable to the interventions being compared in the study. If study participants receive interventions that are not part of the study protocol and could affect the outcomes being assessed, and they receive these interventions differentially, then there is cause for concern because these interventions could bias results. The following scenario is another example of how bias can occur. In a study comparing two different dietary interventions on serum cholesterol, one group had a significantly higher percentage of participants taking statin drugs than the other group. In this situation, it would be impossible to know if a difference in outcome was due to the dietary intervention or the drugs.

Question 11. Outcome measures assessment

What tools or methods were used to measure the outcomes in the study? Were the tools and methods accurate and reliable—for example, have they been validated, or are they objective? This is important as it indicates the confidence you can have in the reported outcomes. Perhaps even more important is ascertaining that outcomes were assessed in the same manner within and between groups. One example of differing methods is self-report of dietary salt intake versus urine testing for sodium content (a more reliable and valid assessment method). Another example is using BP measurements taken by practitioners who use their usual methods versus using BP measurements done by individuals trained in a standard approach. Such an approach may include using the same instrument each time and taking an individual's BP multiple times. In each of these cases, the answer to this assessment question would be "no" for the former scenario and "yes" for the latter. In addition, a study in which an intervention group was seen more frequently than the control group, enabling more opportunities to report clinical events, would not be considered reliable and valid.

Question 12. Power calculation

Generally, a study's methods section will address the sample size needed to detect differences in primary outcomes. The current standard is at least 80 percent power to detect a clinically relevant difference in an outcome using a two-sided alpha of 0.05. Often, however, older studies will not report on power.

Question 13. Prespecified outcomes

Investigators should prespecify outcomes reported in a study for hypothesis testing—which is the reason for conducting an RCT. Without prespecified outcomes, the study may be reporting ad hoc analyses, simply looking for differences supporting desired findings. Investigators also should prespecify subgroups being examined. Most RCTs conduct numerous post hoc analyses as a way of exploring findings and generating additional hypotheses. The intent of this question is to give more weight to reports that are not simply exploratory in nature.

Question 14. Intention-to-treat analysis

Intention-to-treat (ITT) means everybody who was randomized is analyzed according to the original group to which they are assigned. This is an extremely important concept because conducting an ITT analysis preserves the whole reason for doing a randomized trial; that is, to compare groups that differ only in the intervention being tested. When the ITT philosophy is not followed, groups being compared may no longer be the same. In this situation, the study would likely be rated poor. However, if an investigator used another type of analysis that could be viewed as valid, this would be explained in the "other" box on the quality assessment form. Some researchers use a completers analysis (an analysis of only the participants who completed the intervention and the study), which introduces significant potential for bias. Characteristics of participants who do not complete the study are unlikely to be the same as those who do. The likely impact of participants withdrawing from a study treatment must be considered carefully. ITT analysis provides a more conservative (potentially less biased) estimate of effectiveness.

General Guidance for Determining the Overall Quality Rating of Controlled Intervention Studies

The questions on the assessment tool were designed to help reviewers focus on the key concepts for evaluating a study's internal validity. They are not intended to create a list that is simply tallied up to arrive at a summary judgment of quality.

Internal validity is the extent to which the results (effects) reported in a study can truly be attributed to the intervention being evaluated and not to flaws in the design or conduct of the study—in other words, the ability for the study to make causal conclusions

Supplementary S1: Quality Assessment Checklist and Guidelines

about the effects of the intervention being tested. Such flaws can increase the risk of bias. Critical appraisal involves considering the risk of potential for allocation bias, measurement bias, or confounding (the mixture of exposures that one cannot tease out from each other). Examples of confounding include co-interventions, differences at baseline in patient characteristics, and other issues addressed in the questions above. High risk of bias translates to a rating of poor quality. Low risk of bias translates to a rating of good quality.

Fatal flaws: If a study has a "fatal flaw," then risk of bias is significant, and the study is of poor quality. Examples of fatal flaws in RCTs include high dropout rates, high differential dropout rates, no ITT analysis or other unsuitable statistical analysis (e.g., completers-only analysis).

Generally, when evaluating a study, one will not see a "fatal flaw;" however, one will find some risk of bias. During training, reviewers were instructed to look for the potential for bias in studies by focusing on the concepts underlying the questions in the tool. For any box checked "no," reviewers were told to ask: "What is the potential risk of bias that may be introduced by this flaw?" That is, does this factor cause one to doubt the results that were reported in the study?

NHLBI staff provided reviewers with background reading on critical appraisal, while emphasizing that the best approach to use is to think about the questions in the tool in determining the potential for bias in a study. The staff also emphasized that each study has specific nuances; therefore, reviewers should familiarize themselves with the key concepts.

Reference

National Heart Lung and Blood Institute. (2013). *Study Quality Assessment Tools*.
<https://www.nhlbi.nih.gov/health-topics/study-quality-assessment-tools>