Appendix 2. Checklist for assessing the quality of randomized studies

The assessment primarily concerns the quality of a study, that is, the risk of systematic errors and the risk of conflicts of interest (A). The composite assessment of all the included studies according to GRADE also includes agreement across studies (B), generalizability (C), precision (D), publication bias (E), size of the after careful consideration effect (F), dose-response relationship (G) and the probability of underestimating the effect (H).

Author Year Article number

The alternative “unclear” is used when the information is not available in the text. The alternative “not applicable” is used when the question is irrelevant. Specify under comments.

A. Assessment of the study’s limitations, any systematic errors (bias)	**Yes**	**No**	**Unclear**	**Not applicable**
A1. Selection bias				
a) Was a proper randomization method used?				
b) If the study used some kind of restriction in the randomization process (i.e. blocking, stratification, minimization), are the reasons adequate?				
c) Were the groups composed in an adequately similar way?				
d) If imbalances in baseline variables were corrected for, was that done properly?				
Comments:				
Assessment of risk of selection bias:		Low	Moderate	High
A2. Performance bias				
a) Were study participants blinded?				
b) Were personnel blinded?				
c) Was compliance in the groups acceptable according to reliable documentation?				
d) Were the participants otherwise treated/exposed in the same way apart from the intervention?				
Comments:				
Assessment of risk of performance bias:		Low	Moderate	High

A. Continued Yes No Unclear Not applicable

A3. Detection bias (per outcome measure)

a) Was the outcome measure insensitive to detection bias?
b) Were the persons who evaluated the results blinded to the intervention?
c) Were the persons who evaluated the results impartial?
d) Was the outcome defined appropriately?
e) Was the outcome identified/diagnosed with valid methods of measurement?
f) Was the outcome measured at appropriate points in time?
g) Was the choice of statistical measure appropriate for the reported outcome?
h) Was the analysed population (ITT or PP) appropriate for the study question?

Comments:
Assessment of risk of detection bias Low Moderate High

A4. Attrition bias (per outcome measure)

a) Was the attrition acceptably low in relation to the population's size?
b) Was the attrition acceptably low in relation to the size of the outcome?
c) Was the size of attrition balanced between the groups?
d) Were relevant baseline variables balanced between those who broke participation and those who completed the study?
e) Was statistical handling of the attrition adequate?
f) Were the reasons for attrition analysed?

Comments:
Assessment of risk of attrition bias Low Moderate High

A. Continued **Yes** **No** **Unclear** **Not applicable**

A5. Reporting bias

a) Did the study follow a protocol published in advance?
b) Were the primary/secondary outcome measures specified?
c) Were all outcome measures in the study protocol reported comprehensively?
d) Were adverse effects/complications measured adequately?
e) Were the reported outcome measures only those presented in advance in the study protocol?
f) Were the points of time for analysis presented in advance?

Comments:
Assessment of risk of reporting bias Low Moderate High

A6. Conflicts of interest

a) Based on the authors' reported declaration of interest, is there a low or no risk of the results being influenced by conflicts of interest?
b) Based on information about the funding of the study, is there a low or no risk of the study being influenced by economic interests in the results?
c) Is there a low or no risk of other types of conflicts of interest (e.g. it was the authors who had developed the intervention)?

Comments:
Assessment of risk of conflicts of interest Low Moderate High

Summary of risks of bias (per outcome measure)	Low	Moderate	High
A1. Selection bias			
A2. Performance bias			
A3. Detection bias			
A4. Attrition bias			
A5. Reporting bias			
A6. Conflict of interest bias			

Comments:

Summary assessment of risk of systematic errors (bias)	Low	Moderate	High

Basis for composite assessment according to GRADE

B. Insufficient agreement between studies

This is handled only on the level of synthesis

C. **Assessing the study's generalizability**	**Yes**	**No**	**Partly**	**Not applicable**
a) Do the control group's situation and conditions agree with the situation to which the SBU/HTA report's conclusions refer?				
b) Is the included study population reasonably similar to the population to which the SBU/HTA report's conclusions refer?				
c) Is the intervention relevant to the conditions to which the SBU/HTA report's conclusions refer?				

Comments:

Assessment of insufficiencies in generalizability	None	Some	Large

D. Assessing precision	**Yes**	**No**	**Partly**	**Not applicable**
a) Is the precision acceptable considering the number of included individuals and the number of events (outcomes)?				

Comments:

E. Assessing publication bias

This is handled only on the level of synthesis

F. Assessing the size of effects	**Yes**	**No**	**Partly**	**Not applicable**
a) Was the size of the effect large (i.e. RR<0.5 or >2.0)?				
b) Was the size of the effect very large? (i.e. RR <0.2 or > 5.0)?				

Comments:

G. Assessing dose-response relationship

a) Is there support for a dose-response relationship between exposure and outcome?

Comments:

H. The probability that the effect is underestimated due to confounders

Not applicable to RCTs

Checklist for assessing the quality of randomized studies: explanations

The checklist is intended to provide a systematic basis for assessing the risk that a study's estimate of a given outcome was biased during the course of the research work. This can lead to the outcome being either under- or overestimated compared with a "true" outcome. Even the direction of the outcome can be misjudged.

The checklist is intended to result in a systematic and transparent basis for discussing the size of the risk that estimated outcomes in a study are systematically biased. It does not offer an algorithm for summarizing quality points. Regarding judgment bias (A3) and attrition bias (A4), the assessment needs to be done per outcome measure since the shortcomings in quality may differ between outcome measures.

To be able to use the results for grading the quality of evidence according to GRADE, more information is needed, such as summaries on the level of synthesis, that is, a composite assessment if there is more than one study. In some instances, summaries can only be made on the level of synthesis, for example regarding inconsistency, precision and publication bias.

A. Assessing a study's limitations – systematic errors

A1. Risk of selection bias

Selection bias refers to systematic errors related to how the selection of the subjects of an experiment (study participants) was handled and the how the subjects were divided into intervention and control groups.

A risk of selection bias is present if the intervention and control groups are not sufficiently similar at baseline with respect to known as well as unknown risk and protective factors. If they are not sufficiently similar, the outcome can be due, at least in part, to such differences and thereby give biased results. The randomization should be done in an unpredictable way and the process should not be open to manipulation. This can be achieved, for example, by using a computerised random number generator to allocate the participants and by masking the process with sealed envelopes.

Sometimes the randomization is limited in order to get equally large groups (i.e. block randomization) or to balance the groups regarding participant characteristics that may influence the results (i.e. stratified randomization). This procedure may increase the predictability of a participant being allocated to a particular group. This can happen in particular if the blocks are small or if each stratum contains just a small number of individuals.

A1d. Post hoc adjustment of the outcome, based on differences in known baseline factors, is controversial. It can be a reasonable way of testing the sensibility of a positive outcome, but very valid arguments are needed to justify changing a negative outcome in the primary analysis to a positive outcome.

A2 Risk of performance bias

Performance bias refers to systematic errors that are related to how the participants in the intervention-and control groups were handled in the study.

Risk of performance bias is present if the intervention or the control group was exposed to something other than what the comparison aims to measure, for example another treatment for a particular disease than the approved standard treatment. In that case the outcome of the study can be due, at least in part, to such differences and thereby produce distorted results.

If the aim is to estimate the effect of a given intervention, the control group (placebo or un-treated control) should be exposed to exactly the same things as the intervention group apart from the intervention itself. If not, the reported effects may either over- or underestimate the true effect; this also holds for the direction of the effect. In other words, there will be a risk of performance bias.

The differences may concern wrong intervention, incomplete intervention, interrupted intervention, additional intervention outside the study protocol, etc. The risk of bias can be lower if the personnel and the patient are ignorant of grouping (blinded study) and if there is a structured control of the implementation (i.e. a checklist or a manual).

A2a/b It is desirable that both the patient and the investigator (and outcome assessor, see A3b) are blinded in a study. This is sometimes difficult or impossible in practice. The blinding can also fail due to specific effects or adverse effects of the active treatment, such as dry mouth from treatment with neuroleptics and irregular vaginal bleeding from treatment with oestrogen. In order to reduce the risk of jeopardizing the blinding, it is sometimes possible to supplement the active treatment with drugs that counteract adverse effects. Other factors that can impair blinding are differences in the appearance or taste of tablets, inhalation preparations, etc. A large placebo effect can indicate successful blinding. In some studies the participants are invited to guess whether they received the active or the control treatment.

A2c Compliance is particularly important to check when the effect does not differ significantly between the groups. Insufficient compliance can reduce both the intervention's effect and adverse effects. This is particularly important to check in so-called non-inferiority studies, but less important if the intervention has a significant effect. An exception is if compliance was less good in the group receiving the reference intervention. The latter is possible in a placebo-controlled study if blinding was insufficient, or if the reference intervention had a much higher prevalence of adverse effects.

A3. Risk of detection bias

Detection bias refers to systematic errors that are related to how measurements and analysis of the results are handled in the study.

A risk of detection bias is present if there are differences in how the outcomes were decided in the intervention and the control groups. The study's outcome can then be due to this, at least in part, and the results will be distorted. Detection bias, and thereby the study’s overall quality can vary between outcome measures in one and the same study. Detection bias may therefore have to be evaluated separately for each of a study’s outcome measures.

A3a. The risk of bias increases with the number of subjective features in the evaluation of the outcome. While survival/death is a robust outcome measure; symptom scales and quality-of-life measurements are sensitive to bias and are, in principle, useless in unblinded studies.

A3b. Besides being blinded when evaluating the results of the study, it is also important that the study states that the results were processed before the trial code was broken.

A3c. In randomized studies it is often the investigators who also perform the evaluation. Larger, high-quality studies sometimes have independent committees (DSMB) that evaluate and decide the results.

A3d. Here it is often a question of how composite measures, that is, combined outcome measures, are put together, or how surrogate measures are related to clinical relevance.

A3e. Performing the measurement with a standardized method that has been validated in the population in question reduces the risk of bias.

A3f. The timing of measurements in order to optimize the possibility of discovering a difference in the outcome is particularly important in so-called non-inferiority studies.

A3g. The most common outcome measures for dichotomous variables, such as yes-no variables, are risk ratio (RR), odds ratio (OR), absolute risk reduction/risk difference and number needed to treat (NNT). Hazard ratio (HR) is used to analyse risk over time. For continuous variables, absolute difference in means (mean difference) is usually used; alternatively, the response threshold is defined and outcome is reported as responder rate. All measurements (preferably the difference between groups) should be reported with an appropriate measure of precision, preferably the 95 per cent confidence interval.

A3h. The results can be analysed according to intention to treat (ITT) and/or per protocol (PP). An ITT analysis implies that all persons being randomized are followed in their treatment arm irrespective of whether or not they received the selected treatment. This is most often the preferred method. If the results are calculated in some other way than ITT, there is a risk of the treatment effect being overestimated. The ITT analysis can be complemented with a sensitivity analysis according to the “worst case scenario”, where the worst possible outcome is assigned to missing patients in the group with the best effect and the best possible outcome is assigned to missing patients in the group with the worst outcome. Sometimes (particularly non-inferiority studies) it is important that a PP analysis is also reported; only those who completed the whole study protocol are then included in the analysis.

A4. Attrition bias

Attrition bias refers to systematic errors that are related to how the study has handled attrition, that is, persons who agreed to take part in an investigation but dropped out before it had been completed.

A risk of attrition bias is present when the attrition differs between the intervention and the control group. The outcome of the study can be due, at least in part, to these differences and in that way distort the results. A generally large attrition, differences in attrition and above all differences in the reasons for attrition increase the risk of bias. One can never count on attrition occurring at random. If the composition of persons who dropped out did not differ from those who were still in the study, the situation is better than if there are significant differences between the two groups. Examples given below can serve as *crude* benchmarks:

- Small (<10 %)
- Moderate (10-19 %)
- Large (20-29 %)
- Very large (≥30 %). The study is often judged to have no informational value, which can mean that it should be excluded.

Attrition must also be seen in relation to the size (and difference) of the outcome. The smaller the outcome, the greater the problem even with small attrition rates.

Attrition can differ between points in time of the study as well as between measures of outcome. Attrition grows as time passes. The results of a treatment from the latest visits may therefore be of questionable validity, while results from the first visits may be valid.

A4e. Attrition is analysed using methods for imputation (missing measurements are replaced; for example, last observation carried forward (LOCF), observed cases (OC) or interpolation). It is important that the results with different imputation methods are reported or that a method is used that is least favourable to the outcome (conservative). This may well result in the size of the effect being underestimated. In so-called non-inferiority studies it is better to use an imputation method that favours the outcome, since otherwise you will draw the wrong conclusion about the absence of effect/difference.

A5. Reporting bias

Reporting bias refers to systematic errors that are related to how the study handles the reporting in relation to its protocol.

A study's outcomes can be due, at least partly, to only certain results being reported instead of every result. In that case the outcome can be either over- or underestimated. The direction of the outcome can also be influenced.

A5a. It is not uncommon for studies with negative results to include explanatory or post hoc-analyses in order to find certain subgroups of patients who may benefit from the treatment. These analyses can serve to generate hypotheses, but the conclusions from a *negative* study must never be based on such analyses. When, however, a study shows a *positive outcome* for the primary outcome measure, subgroup analyses are valuable for assessing the results' generalizability.

A5c/d. Even if the reported outcome measures are reasonable, pre-defined and adequately reported, other important outcome measures may have been excluded. This often applies to outcome measures for assessing adverse effects/risks.

A5f. It is important that all the analyses feature in the protocol (and the statistical plan). It is also important that the study states whether the reported analysis is the final analysis or a pre-planned interim analysis. Ad hoc interim analysis is of course problematic, particularly in open studies where the analysis can be suspected of being data driven.

A6. Conflict of interest bias (other considerations)

If the author(s) of the study stand to gain from a certain result, this can lead to over- or underestimation of the effect in the direction that favours the authors.

Summary assessment

To be able to decide on the quality of the evidence for an outcome according to GRADE, all the above types of risk of bias must be weighted together and included in a composite effect of an outcome from one or more studies. This can preferably be discussed by a group of experts.

B. Insufficient agreement between studies (heterogeneity)

To be handled on the level of synthesis.

C. Insufficient generalizability (indirectness of evidence)

Generalizability refers to the possibility of applying a study's design, discussion and results to the conditions for the SBU/HTA report.

Problems with generalizability arise if the population, control alternatives, or outcome measures differ from those specified for the SBU/HTA report, .The outcome presented in the study may then differ, at least in part, from the "true" outcome concerning how the population, intervention, control alternatives and outcome measures have been specified in the review. The outcome can be under- or over-estimated, and the direction of the outcome can be affected.

That the study population corresponds to the population from which the SBU/HTA report intends to draw conclusions is much more important than that the study population is not representative of the aim of the particular study (e.g. depending on attrition before randomization).

The included studies must be treated as a whole to enable the dimension indirectness to be considered when grading the quality of evidence according to GRADE.

D. Lack of precision (imprecision)

There are two aspects of precision to consider. Firstly, if the aim is to test whether the intervention is better than the control alternative, it is sufficient to study whether the confidence interval straddles the line of "no difference" ("1" for binary outcomes and "0" for continuous outcomes). If this line is straddled, the precision is insufficient. In Figure B2.1, the results for Superior, Non-inferior B and Inferior have good precision in this respect. Secondly, if the aim is to test whether the intervention is worse than the control treatment (often regarding adverse effects), a pre-specified clinically-defined limit to how much worse the intervention can be without being a problem (suggested appreciable harm) is also required. If the confidence interval does not straddle this limit, the precision is good and one can conclude that the intervention is not worse than the control treatment. In Figure B2.1 the limit was set to 1.25. Three examples of results illustrating good precision are Superior, Non-inferior A and Non-inferior B. Examples of poor precision are Imprecise A and Imprecise B. Note that the quality of the data is important when judging the precision in non-inferiority outcomes. For example, poor reporting of adverse effects can lead to the results looking equally good in the two treatment arms.

Figure B2.1. Illustration of different tests using a forest plot.

If there are several studies suitable for weighting together, a composite confidence interval should be considered.

E. *Publication bias*

To be handled only on the level of synthesis.

F. *Size of effects*

Primarily handled on the level of synthesis. If the quality of included studies has been downgraded, after careful consideration the size of the effect can be upgraded.

G. *Dose-response relationship*

Weighted together on the level of synthesis. It can be practical to note the results of single studies in the checklist.

H. *Probability of an underestimated effect due to confounders*

Not applicable to RCT s.