Appendix 3. Checklist for assessing the quality of observational studies

The assessment of a study primarily concerns its quality, 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 between studies (B), generalizability (C), precision (D), publication bias (E), size of the 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) Were the groups recruited in a similar way?				
b) Was the composition of the groups sufficiently similar at the start of the study?				
c) Were any corrections to imbalances between groups with different exposures/treatments made properly in the statistical analysis?				
Comments:				
Assessment of risk of selection bias:		Low	Moderate	High
A2. Performance bias				
a) Were the groups' conditions (apart from the studied treatment or exposure) sufficiently similar during the treatment/exposure?				
b) Was the groups' compliance with the treatment/exposure acceptable?				
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 study participants' exposure status?
c) Were the persons who evaluated the results *impartial*?
d) Was the outcome defined appropriately?
e) Was the outcome measured using standardized/ defined measurement methods?
f) Was the outcome measured adequately, using validated measurement methods?
g) Were variations in exposure over time included in the analysis?
h) Was the outcome measured at optimal points in time?
i) Was observer agreement acceptable?
j) Did the study use a suitable statistical measure for the reported effect/association?

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 size of the population?
b) Was attrition the same size in each of the groups?
c) Was the distribution of relevant baseline variables uniform in attritions in the intervention and control groups, alternatively in different exposure groups?
d) Was the statistical handling of the attrition adequate?

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 that had been established in advance?				
b) Were the outcome measures relevant?				
c) Were adverse effects/complications measured adequately?				
d) Were the points of time for analysis relevant?				

Comments:

Assessment of risk of reporting bias Low Moderate High

A6. Conflicts of interest				
a) Based on the authors' reported declarations of interest, is there a low or no risk that the results have been influenced by conflicts of interest?				
b) Based on information about the funding of the study, is there a low or no risk that the study has been 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

Weighted summary of risk 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:

Composite assessment of the risk of systematic error (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 context and the control group's conditions agree with the situation to which the conclusions in the SBU/HTA report refer?				
b) Is the included study population reasonably similar to the population to which the conclusions in the SBU/HTA report refer?				
c) Is the intervention relevant to the conditions to which the conclusions in the SBU/HTA report 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

Handled on the level of synthesis

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

Occasionally the quality of evidence can be up-graded if it is highly likely that the effect is underestimated.

a) Is there strong support for the possibility that confounders which could not be considered in the study would have strengthened the association?

Comments:

Checklist for assessing the quality of observational studies: explanations

The check list is primarily intended to be used for assessing the quality of prospective cohort studies (part A). It can also be used with certain additions/adaptations for retrospective cohort studies with historical controls, retrospective cases series, cross-sectional studies or other non-randomized types of study when it is appropriate to include such studies.

The checklist is intended to provide a systematic basis and support for assessing the risk that a study's estimate of a given outcome has been 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 allotted to intervention and control groups.

A risk of selection bias can arise if the intervention and control groups are not sufficiently similar at baseline in terms of known and unknown risk and protective factors so as not to distort the results. Some important confounders are age, gender, underlying disease history and co-morbidity. Another such factor is socio-economy, which is probably the most important risk factor for morbidity and untimely death. Corrections for known confounders can be made with statistical methods such as matching, stratification, multivariate regression analysis and propensity score methodologies (see A1c).

There is also a high risk of selection bias if the measure is particularly suitable for certain experimental subjects who are particularly likely to respond well to the measure.

A1a. Is the comparison group clearly defined? Were the groups to be compared recruited in reasonably similar ways so that the results were not distorted? Was the comparison group recruited from the general population or from a restricted selection. If the comparison group was a historical control group, the results need to be assessed with special care. An important issue is whether the intervention and control groups were recruited with the same method.

A1b. Data that can reveal substantial differences between the groups are often found in an introductory table or as background data (baseline characteristics).

A1c. Methods that can be used in this connection are matching/restriction, stratified analysis, multivariate modelling analysis (e.g. regression analysis) or propensity score methodology.

Since observational studies (2+) in the GRADE system are assumed from the start to have a higher risk of selection bias than randomized studies, the risks of selection bias must be considered very high in this section.

A2. Risk of performance bias

Performance bias refers to systematic errors related to how the study handled persons in the intervention group and the comparison group.

A risk of performance bias is present when the intervention or control group is exposed to something other than what the comparison aims to measure. The measured effect may then be caused, at least in part, by such differences and thereby distort the results. Differences may, for example, concern wrong treatment, incomplete treatment, interrupted treatment or additional treatment outside the study protocol. Structured control of the implementation (e.g. a checklist or a manual) can reduce the risk of systematic errors.

A2a. If the study aims to estimate the effect of a certain intervention/risk factor (possibly in relation to an alternative intervention), the control group should be exposed to exactly the same thing as the intervention group apart from the intervention itself. Otherwise the effect can be over- or underestimated. This also applies to the direction of the effect, that is, there is a risk of performance bias. For example, are there socio-economic differences between the intervention and control groups? The risk is particularly high as regards preventive and symptom-alleviating measures that well-informed individuals can obtain, resulting in the effect/risk being underestimated. If the groups are exposed to different factors that affect the outcome in a similar way, this may also reduce the possibility of the study detecting or excluding an effect or a risk association.

A2b. Controlling compliance with an intervention or an exposure is fundamental for the reliability of the results. This is particularly important when the result indicates a lack of effect/association. This may be due to a lack of exposure to the risk factor or intervention. The extreme form of low compliance is withdrawal from an intervention or exposure, that is, the experimental subject interrupts the treatment or terminates the exposure (without having provided the studied outcome) but does not necessarily interrupt the follow-up (=attrition, see A4). It is important to check:

a) the total number of interruptions
b) the difference in interruptions between groups
c) differences in the reasons for interruption between groups

A3. Risk of detection bias

Detection bias refers to systematic errors related to how the study handles measurements and analysis of results. A risk of detection bias is present if there are differences in how the outcomes were decided in the intervention and the control group. The study's outcome can then be due to this, at least in part, which will distort the results. Detection bias, and thereby a study's overall quality, can differ 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.

The answers to certain questions can make other questions less or not at all relevant. The relevance of questions on blinding (A3b), for example, depends on the robustness of the outcome variable (A3a).

A3a. The risk of bias increases with the number of subjective features in the evaluation of the outcome. 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. The risk of systematic error can be greater if those who measure the outcome (pathologist, radiologist, psychologist) or evaluate the results of the measurement ("researcher") know which experimental subjects received a certain treatment/exposure.

A3c. The risk of systematic error also increases if the personnel taking part in the intervention or the conduct of the study also assess the outcome. This impartiality is of little importance if blinding is used.

A3d. Here it is often a question of how so-called composite measures, that is, combined outcome measures, are assembled, or how different surrogate measures are connected to clinical relevance. If the outcome is negative, it is important that the selected effect measure is sufficiently sensitive and the confidence interval sufficiently narrow to be able to exclude the possibility that the size of the effect is clinically relevant.

A3e/f. The risk of bias is lower if the measurement is performed with a standardized or defined method that has been validated in the population in question.

A3g. The possibility of detecting (as well as excluding) effects/associations increases if the exposure is assessed at repeated (optimal) points in time during the study.

A3h. Choosing the wrong point in time for measuring can lead to the results being underestimated. This is particularly important for non-inferiority ("not worse than") studies or a conclusion that there is no effect.

A3i. Observer variability can be a weakness when the outcome is recorded. An example is when radiographs or cytological tests are evaluated by more than one observer. In such cases, the agreement between all or most of the observers should be reported. Depending on the scale used, this can be done with kappa-agreement or intra-class correlation coefficient (ICC).

A3j. The most common outcome measures used 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) and standardized mean difference (Cohen's d, Hedge's g) are usually used or alternatively the threshold for response and outcome is defined and reported as responder rate. When using such dichotomization of continuous variables, it is important that the interval threshold(s) is properly motivated or is currently accepted.

All measurements (preferably the difference between groups) should be reported with the appropriate measure of precision, preferably the 95 per cent confidence interval. Assess whether confidence intervals or other relevant measures are adequately reported; if not, whether not reporting them is motivated. That can be the case, for example, with total investigations of large data materials.

A4. Attrition bias

Attrition bias refers to systematic errors related to how the study has handled attrition, that is, persons who agreed to take part in an investigation but dropped out before it was completed (loss to follow-up).

A risk of attrition bias is present when attrition differs between the intervention and control groups. A study's outcome 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. An assessment of attrition concerns attrition after inclusion in the study. One can never count on attrition occurring at random.

Large attrition increases the risk of the results being influenced by systematic errors. Attrition can differ between points in time and between outcomes. Attrition therefore has to be analysed separately for each outcome. A long follow-up time may be a reason for accepting a larger attrition.

Attrition can differ between points in time in the study and 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.

A4a. As a benchmark value for studies on pharmaceuticals, the risk is small if the attrition is less than 10 per cent, moderate if the attrition is between 10 and 19 per cent and large of the attrition is between 20 and 29 per cent. An attrition of 30 per cent or more makes the informative value doubtful and the study may have to be excluded. Note that there may be other benchmarks for other types of study. The attrition must also be judged in relation to the size (and difference) of the outcome. The smaller the outcome the greater the problem, even with small attrition rates.

A4c. Differences in attrition between baseline variables in the intervention and control groups or between groups with different exposures to risk factors are serious because they can distort the outcome, particularly if the differences concern baseline factors directly related to the outcome (e.g. stage of disease for survival outcomes).

A4d. If the composition of persons in the attrition differs from those still in the study, this can affect the possibility of the study discovering relevant effects and generalizability (e.g. patients with progressing disease who cannot fill in questionnaires about quality-of-life).

A4e. An analysis of attrition includes various methods for imputation (missing measurements are replaced, for example, by last observation carried forward (LOCF), observed cases (OC) or interpolations). It is important that the results with different imputation methods are reported or that the study uses the method 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 because otherwise the wrong conclusion will be drawn about the absence of effect/difference.

A5. Reporting bias

Reporting bias refers to systematic errors 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. The outcome can then either be over- or underestimated. The direction of the outcome can also be influenced.

A5a. Access to the study protocol is of great value for assessing the importance of the reported results since 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, alternatively that associations between subgroups are identified. These analyses can serve to generate hypotheses but the conclusions from a *negative* study must never be based on such subgroup analyses. When, however, a study shows a *positive outcome* for the primary outcome measure, subgroup analyses are valuable for assessing the results' generalizability.

A5b. It is important to clarify which outcomes are measured, analysed and reported.. Outcomes that are measured or analysed but not reported and therefore not taken into account in the statistical analysis may lead to the importance of the intervention/association being misinterpreted.

A5c. Insufficient reporting of an intervention's risks can lead to its suitability being overestimated (benefit/risk).

A5d. Effects may have been measured repeatedly and it is important that the analyses are limited to those given in the protocol (and the statistical plan admits). It is also important that the study states whether the reported analysis is the final one 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. Even "lege artis" interim analyses run a risk of overestimating an intervention's effects. In the case of studies that do not show any effect of an intervention, it is important that the point in time for the analysis is optimal for detecting an effect.

A6. Conflict of interest bias (other considerations)

If the author(s) of the study stand to gain from certain results, this can lead to over- or underestimation of the effect in the direction that favours the authors. For example, it can be problematic if it was the authors who have developed the intervention in question.

Summary assessment

In order to assess the composite evidence using GRADE, the factors mentioned below must also be weighted together in a final judgment.

B. *Insufficient agreement between studies (heterogeneity)*

If possible, this assessment is done in the form of meta-analyses or similar analyses. It is excluded when a single study is assessed.

C. *Insufficient generalizability (indirectness of evidence)*

Generalizability refers to the possibility of using a study's design, discussion and results under the conditions that apply for SBU/HTA reports.

Problems with generalizability arise if the population, intervention, control alternatives, or outcome measures in a study differ from those specified for SBU/HTA reports as adequate for Swedish conditions and thereby for the question at issue. The outcome presented in the study may 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 effect can be under- or overestimated in relation to Swedish conditions.

It is important that the study population corresponds to the population for the SBU/HTA report.

In order to consider indirectness when grading the quality of evidence according to GRADE it is necessary to have a composite weighted outcome and treat the included studies as a whole.

D. *Lack of precision (imprecision)*

Two aspects of precision are considered here. Firstly, if the aim is to test whether the intervention is better than the control alternative, it is sufficient to consider whether the confidence interval straddles the line of "no difference" ("1" for binary outcomes, "0" for continuous outcomes). A confidence interval that straddles this line indicates insufficient precision. In Figure B3.1, 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 B3.1 the limit is set to 1.25. Three examples of such results 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 assessing 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 B3.1. Illustration of tests using a forest plot.

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

E. *Publication bias*

Weighted together on the level of synthesis.

F. *Size of effects*

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

G. *Dose-response relationship*

The final assessment is handled on the level of synthesis.

H. *Probability of an under-estimated effect due to confounders*

Weighted together on the level of synthesis. It can be practical to note the results of single studies in the checklist. Occasionally the strength of evidence can be adjusted and up-graded of it is highly likely that the studies underestimate the effect. This can happen if confounders for which the studies could not control indicate that the effect is underestimated.