GUIDELINE

Internal validity of non-randomised studies (NRS) on interventions

July 2015
The primary objective of EUnetHTA JA2 WP 7 methodology guidelines is to focus on methodological challenges that are encountered by HTA assessors while performing relative effectiveness assessments of pharmaceuticals or non-pharmaceutical health technologies.

As such the guideline represents a consolidated view of non-binding recommendations of EUnetHTA network members and in no case an official opinion of the participating institutions or individuals.

Disclaimer: EUnetHTA Joint Action 2 is supported by a grant from the European Commission. The sole responsibility for the content of this document lies with the authors and neither the European Commission nor EUnetHTA are responsible for any use that may be made of the information contained therein.
This guideline has been developed by
IQWiG (Institute for Quality and Efficiency in Health Care), Germany

With assistance from draft group members from
NOKC (Norwegian Knowledge Centre for the Health Services), Norway
SNHTA (Swiss Network for Health Technology Assessment), Switzerland

The guideline was also reviewed and validated by a group of dedicated reviewers from
AIFA – IT
ZIN – NL
NETSCC – UK
SBU – SE
NICE – UK
Table of contents

Acronyms - Abbreviations .................................................................................................................. 5
Summary and table with main recommendations ............................................................................... 6
1. Introduction ...................................................................................................................................... 8
   1.1. Definitions of central terms and concepts .............................................................................. 8
   1.2. Problem statement .................................................................................................................... 9
   1.3. Objective(s) and scope of the guideline ................................................................................. 10
   1.4. Related EUnetHTA documents .............................................................................................. 11
2. Analysis and discussion of the methodological issue .................................................................... 12
   2.1. Key criteria for RoB tools ....................................................................................................... 12
   2.2. Systematic review of the literature on RoB tools .................................................................. 13
   2.3. Evaluation of existing RoB tools ............................................................................................. 13
   2.4. Reliability and ease-of-use of RoB tools ............................................................................... 16
   2.5. Summarizing the results of RoB assessments ....................................................................... 16
   2.6. Presenting the results of RoB assessments ............................................................................ 17
   2.7. Summary of the results ........................................................................................................... 18
3. Conclusion and main recommendations ....................................................................................... 19
   Annexe 1. Bibliography .................................................................................................................. 20
   Annexe 2. Documentation of literature search ............................................................................... 24
   Annexe 3. Example table on how to present RoB assessment results ............................................ 33
Acronyms - Abbreviations

ACROBAT – A Cochrane Risk of Bias Assessment Tool
RoBANS – Risk of Bias Assessment Tool for Non-randomized Studies
CONSORT – Consolidated Standards of Reporting Trials
EUnetHTA – European network for Health Technology Assessment
GRADE – Grading of Recommendations, Assessment, Development and Evaluation
HTA – Health technology assessment
IQWiG – Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (Institute for Quality and Efficiency in Health Care)
NRS – Non-randomised study
PRISMA – Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT – Randomised controlled trial
RoB – Risk of Bias
STROBE – Strengthening the Reporting of Observational Studies in Epidemiology
WP – Work package
Summary and table with main recommendations

This guideline is intended to provide recommendations on the assessment of the internal validity of non-randomised studies (NRS) used for the evaluation of effects of interventions. The inclusion of NRS in a systematic review conducted as part of an HTA may be useful in specific circumstances, but leads to several challenges in terms of internal validity assessment. The aim of this guideline was to recommend tools or checklists that are suitable for assessing risk of bias (RoB) in NRS evidence.

RoB tools were identified from previous systematic reviews and own systematic literature searches. Key criteria, such as coverage of relevant bias domains, were used to evaluate the tools. In addition, tools were required to be free of items on reporting quality and external validity (or applicability). Literature findings concerning reliability and ease-of-use were used as additional criteria.

A total of 11 tools were identified and assessed in detail. Two tools emerged as the currently best instruments for assessing RoB in NRS: ACROBAT-NRSI (A Cochrane Risk of Bias Assessment Tool) and RoBANS (Risk of Bias Assessment Tool for Nonrandomised Studies). As both tools have been developed only very recently, data on reliability are sparse, but it is clear that adequate training is required before assessing NRS evidence. Because ACROBAT-NRSI offers endpoint-specific assessments, requires a summary rating and comes with detailed instructions and documentation guides, this tool is recommended as primary RoB tool for assessment of NRS.

Recommendations

<table>
<thead>
<tr>
<th>Recommendations</th>
<th>The recommendation is based on arguments presented in the following parts of the guideline text</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st recommendation:</td>
<td>1.2</td>
</tr>
<tr>
<td>As the inclusion of non-randomised studies (NRS) in an HTA report requires large efforts (but often fails to increase the validity of the report’s conclusion), the decision to do so should be made only after careful consideration of all advantages and disadvantages.</td>
<td></td>
</tr>
<tr>
<td>2nd recommendation:</td>
<td>2.1</td>
</tr>
<tr>
<td>Internal validity (or risk of bias) should be assessed separately from quality of reporting and external validity (or applicability).</td>
<td></td>
</tr>
<tr>
<td>3rd recommendation:</td>
<td>2.1</td>
</tr>
<tr>
<td>Assessment of risk of bias (RoB) covers at least 5 different types of bias: selection bias (including bias due to confounding), performance bias, detection bias, attrition bias, and reporting bias.</td>
<td></td>
</tr>
</tbody>
</table>

The recommendation is based on arguments presented in the following parts of the guideline text.
<table>
<thead>
<tr>
<th>Recommendation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>4th recommendation:</td>
<td>At present, ACROBAT-NRSI (A Cochrane Risk of Bias Assessment Tool) should be used for the RoB assessment of NRS.</td>
</tr>
<tr>
<td>5th recommendation:</td>
<td>Adequate methodological and clinical knowledge is required for valid and reliable RoB assessment in NRS, because a full understanding of both bias mechanisms and possible confounders is necessary. In addition, clear and consistent decision rules should be agreed on to achieve acceptable reproducibility.</td>
</tr>
<tr>
<td>6th recommendation:</td>
<td>RoB assessment requires that NRS evidence is first subdivided into cohort and case-control studies. Registry analyses usually fall into the category of cohort studies.</td>
</tr>
<tr>
<td>7th recommendation:</td>
<td>In HTA reports, RoB assessment of NRS should be described in meticulous detail in order to enable readers to understand the process and the results.</td>
</tr>
</tbody>
</table>
1. Introduction

1.1. Definitions of central terms and concepts

- **Applicability**, also known as external validity, generalisability, or transposability, is the extent to which the effects observed in clinical studies are likely to reflect the expected results when a specific intervention is applied to the population of interest.

- **Attrition bias**: is caused by missing outcome data. Possible reasons for missing outcome data include loss to follow-up, incomplete data collection, and exclusions of study participants.

- **Bias**: a systematic error in an estimate or an inference. Because the results of a study may in fact be unbiased despite a methodological flaw, it is appropriate to consider risk of bias (RoB).

- **Case-control study**: a study design which identifies patients who have developed an outcome and compares their past exposure (including interventions) with that of controls who do not have the outcome.

- **Cohort study**: a study design where a sample of persons with and without an exposure (including interventions) is followed over time in order to compare the incidence of an outcome between exposed and unexposed persons.

- **Confounding**: Confounders are pre-intervention variables that are associated with the intervention and causally related to the outcome of interest. Bias due to confounding predominates in observational studies.

- **Detection bias**: occurs when outcome measurement is affected systematically by intervention status. This bias may be present, if outcome assessors are aware of intervention status, if different methods of outcome assessment are used in the different interventions groups, or if intervention status affects measurement errors.

- **Internal validity**: the extent to which the (treatment) difference observed in a trial is likely to reflect the ‘true’ effect within the trial (or in the trial population) by considering methodological criteria.

- **Non-randomised study**: a study design that lacks randomised allocation of interventions. Non-randomised (or observational) studies can be grouped into cohort, case-control, and non-comparative studies.

- **Performance bias**: arises due to departures from the intended interventions. This type of bias can be caused by co-interventions, treatment switches, contamination, and other failures to implement the intervention as intended (e.g. non-adherence).

- **Reporting bias**: is defined as the selective reporting of study results depending on the nature or direction of results. Reporting bias may be present both on the study level (e.g. non-publication of complete study) and on the outcome level (e.g. non-publication of outcomes within published studies). These two subtypes of bias have been called publication bias and outcome reporting bias.

- **Selection bias**: occurs when selection of study participants into the study is related to an effect of the intervention or a cause of the intervention and an effect of the
outcome or a cause of the outcome. The term selection bias should not be used to describe bias due to confounding or applicability (i.e. external validity).

1.2. Problem statement
Although randomised controlled trials (RCTs) provide the most robust evidence, other types of studies may provide additional information on the relative efficacy or effectiveness of medical interventions (1). The debate concerning the relative advantages of randomised and non-randomised studies has been going on for decades (2-25). Both types of research designs should probably not be seen as opposing each other but as complementary. No general recommendation can be made as to which alternative is preferable, because the decision depends on topic-specific circumstances, regulatory context, resources and time expenditure.

Possible reasons favouring the inclusion of non-randomised studies (NRS) include:

- The research question cannot (or only with the greatest difficulty) be answered in RCTs. This may be the case because of organizational reasons (e.g. in public health interventions) or epidemiologic circumstances (e.g. very rare diseases).

- The research question can probably be answered with NRS evidence, because very large effects are likely (or at least possible).

- There is an external need to offer a ‘best guess’ rather than no answer at all. Such a situation may be present early in the life cycle of a new intervention or when HTA is used to make only a temporary decision which is followed by an early reassessment (e.g. in a coverage with evidence development [CED] model).

Possible reasons against inclusion of non-randomised studies (NRS) include:

- The HTA report aims at providing a highly reliable result. The inclusion of NRS as the sole information source will very often prevent the results from being ‘definitive’.

- There is an external need to complete the HTA report within a short time period. As indexing of studies in electronic databases and reporting of study details is less complete for NRS than for RCTs, HTA-associated workload increases when NRS are included.

- The inclusion of NRS evidence might mislead researchers into the false belief that RCTs are not worthwhile to perform. Thus, HTA might act as a barrier in finding out the ‘true’ effect of an intervention.

The reasons favouring the inclusion of NRS have considerably less weight, if it is clear that RCTs (of adequate quality and sample size) exist. In the following, it therefore was assumed that HTA reports are more likely to include NRS as the sole rather than an additional source of information on effectiveness and safety.

The inclusion of NRS leads to specific challenges in terms of internal validity assessment (26-28). Classifying the design of a given study can also be difficult, because methodological descriptors in a scientific article may be wrong or missing (29, 30). Due to the large variety of NRS designs and their varying susceptibility to different biases, it is complex to perform a uniform evaluation of RoB for this type of evidence (31-34).
1.3. Objective(s) and scope of the guideline

This guideline is intended to provide recommendations on the assessment of the internal validity of NRS used for the evaluation of effects of interventions. For this scope two main questions originally came into focus:

- How to classify NRS evidence according to study design and
- how to best assess risk of bias (RoB) of specific NRS types.

The classification of study designs, however, was considered relevant for the purposes of this guideline only if it supported the assessment of internal validity in the context of an RoB instrument. It was preferred to recommend only one tool or checklist that – partly or fully – is applicable to different types of studies (ideally including also RCTs). If this would have not been possible, separate tools for the assessment of the most important study designs would have been proposed.

The classification of study designs (i.e. the first question of the guideline) cannot be answered on the basis of empirical arguments alone but requires a conceptual model found in epidemiology. Common labels of study designs such as prospective cohort study or case-control study are ambiguous and require clear definition before being used as eligibility criteria or indicators of RoB.

The guideline did not aim at specifying exactly when to include NRS evidence, as the choice between RCT and NRS evidence is an often debated and multifaceted topic (see section 1.2). Thus, it appears appropriate to address it in a separate guideline. However, some general comments on this issue were deemed necessary (see previous section). Furthermore, the scope of the present guideline was restricted to assessments of intervention effects (therapeutic, preventive, screening or diagnostic interventions), thereby excluding all studies on diagnostic or prognostic test accuracy, aetiology or epidemiology. Finally, assessments of external validity or applicability were considered outside the scope of the present guideline, as this topic is addressed in another EunetHTA guideline (35).

When assessing the RoB of NRS (i.e. the second aim of the guideline), it is worthwhile to keep in mind that NRS may be affected by exactly the same forms of bias that can occur also in randomised studies (e.g. attrition bias or information bias). Therefore, tools for assessing RoB could include partly the same items for RCT and NRS. Thus, this guideline on NRS built on the existing EUnetHTA guideline on the internal validity of RCTs (36). Nevertheless, because of the non-random allocation of research participants to groups, selection bias and confounding are likely to be introduced in NRS. When evaluating NRS data, assessing these issues therefore is extremely important and should most likely be focused on the different methods of statistical adjustments used in such studies (e.g. matched-pair analysis, multivariate regression models, g-estimation or propensity score matching).

Inclusion of non-comparative studies such as case series poses additional difficulties to researchers, because the lack of between-group comparisons precludes assessment of relative effectiveness. Given the lower importance of non-comparative studies for relative effectiveness assessment (REA) and HTA, it was deemed less important to propose any formal tool for assessing RoB of bias in these types of studies. In the assessment of safety, however, non-comparative studies may play a greater role (37).
An ideal assessment instrument would be valid, reliable, easy-to-use and widely applicable. The aim of the present guideline was to systematically identify such an ideal RoB assessment method.

1.4. Related EUnetHTA documents
Other EUnetHTA methodology guidelines should be consulted when assessing the internal validity of RCTs (36) or when assessing external validity or applicability (35). Furthermore, the issue of diagnostic test accuracy and personalised medicine is or will be addressed in other EUnetHTA guidelines or methodological papers.
2. Analysis and discussion of the methodological issue

2.1. Key criteria for RoB tools

Before a systematic search for RoB tools can be conducted, it is essential to define key criteria that high-quality RoB tools should fulfill. The following criteria were applied:

a) Suitability of RoB tool for cohort or case-control studies (at least one required),

b) Suitability of RoB tool for all fields of medicine (i.e. no disease-specific or ad hoc tools)

c) Assessment of internal validity with no inclusion of items on quality of reporting or applicability (i.e. external validity)

d) Coverage of all main types of bias (at least 5 different domains)

These four criteria were considered mandatory and are shortly explained in the following paragraphs, before describing two additional non-mandatory criteria.

For being eligible under criterion a), an RoB tool was required to be suitable for assessing cohort studies, case-control studies, or both. It was not required that RoB tools were also suitable for assessment of case series or RCTs. Furthermore, the project excluded tools that are used to assess a body of evidence rather than individual studies. Thus, concepts such as GRADE (Grading of Recommendations, Assessment, Development and Evaluation) were not eligible.

Criterion b) is required, because the guideline aims at supporting HTA in all fields of medicine (e.g. pharmaceuticals, medical devices, screening, or surgical interventions).

With regard to criterion c), any assessment of published research should avoid mixing up quality of reporting and quality of research (i.e. RoB). Reporting standards, such as CONSORT, PRISMA, or STROBE, are well-established nowadays and have greatly increased transparency in clinical research. Nevertheless, the validity of study results cannot be improved by clear reporting. It is therefore not important for HTA whether an article includes an indicative title, a well-grounded hypothesis or a balanced discussion. Thus, RoB tools that included items on quality of reporting were excluded.

As also set out in criterion c), RoB tools should not mix internal with external validity (or applicability). While internal validity concerns the quality of the design, performance and analysis of a study, external validity relates to how well study results can be generalised to other patients or settings. Because external validity is dependent on the clinical setting, it is in the eye of the beholder and can never be proven on a ubiquitous level. Therefore, assessment of RoB does not include external validity, and any RoB tool containing items on external validity was deemed less suitable for recommendation. Removing inappropriate items of a RoB tool in order to fulfill criterion c) was not considered an option, because any modification of an existing tool would mean creation of a new one, which in turn would require extensive evaluation and piloting.

As described in the EUnetHTA guideline on RCT assessment (36), at least 5 different types of bias are important: selection bias, performance bias, detection bias, attrition bias, and reporting bias. In the context of NRS, selection bias is probably the most important problem, as not only sampling of research participants but also assignment to interventions can be unequal. This latter type of bias is often termed confounding bias. The former type of bias, which relates to the selection of participants into the study, especially affects
those research designs where selection of participants differs between groups, e.g. case-control studies or studies with a historical control group. Because of these considerations and in line with the previous review by Deeks et al. (31), criterion d) required for eligibility that RoB tools addressed at least 5 different domains of bias, including selection bias, confounding, or both.

As mentioned before, additional criteria for choosing an RoB tool are ease-of-use and acceptance among systematic reviewers. These criteria were considered as additional non-mandatory criteria.

2.2. Systematic review of the literature on RoB tools

A systematic review of the literature was performed to identify tools for assessing RoB in NRS. Because two previous systematic reviews with the same focus had performed literature searches in 1999 and 2005 (31, 38), the current searches were limited to articles published from 2005 onwards. Details of the literature search can be found in Annexe 2. The bibliographic search identified one new RoB tool, called RoBANS (Risk of Bias Assessment Tool for Non-randomised Studies) (34). Out of a total of 30 tools, RoBANS was the only one which fulfilled the key criteria.

In addition to bibliographic searches, the Non-Randomised Studies Methods Group of the Cochrane Collaboration was contacted in order to obtain information with regard to a new RoB tool, which was then under development. The Cochrane Group provided the prefinal version of the new tool, called ACROBAT (A Cochrane Risk of Bias Assessment Tool). The authors of this guideline participated in piloting this new tool.

In the systematic review by Deeks et al. (31), only 6 assessment tools (from a total of 193 tools) were “judged to be potentially useful” (39-44). The systematic review of Sanderson et al. (38) failed to identify “a single obvious candidate tool” for RoB assessment of NRS. Through contacts with stakeholders and cross-referencing, an additional 4 eligible tools were identified (45-48). Therefore, a total of 11 tools required more detailed evaluation:

- ACROBAT-NRSI (A Cochrane Risk of Bias Assessment Tool) (48)
- the Berger/ISPOR questionnaire (46)
- the Cowley checklist (40)
- the Downs-Black checklist (41)
- EPHPP (Effective Public Health Practice Project Quality Assessment Tool) (43)
- the GRACE checklist (Good ReseArch for Comparative Effectiveness) (47)
- MINORS (Methodological Index for Non-randomised Studies) (45)
- NOS (Newcastle-Ottawa Scale) (44)
- the Reisch-Tyson checklist (39)
- RoBANS (Risk of Bias Assessment Tool for Non-randomised Studies) (34)
- TFCPS (Task Force on Community Preventive Services) (42)

2.3. Evaluation of existing RoB tools

The following text gives a brief description of each RoB checklist together with a discussion of its main problems. A summary of how the checklists cover the different bias domains can be found in Table 1 below.

ACROBAT-NRSI was developed in 2014 by the Non-Randomised Studies Methods Group (NRS MG) of the Cochrane Collaboration and other partners. The total number of items varies between 22 and 29, as some items are to be filled in only in specific cases. The instrument requires a RoB judgment (low, moderate, serious, or critical) in each of the 7 do-
mains and for the overall assessment. Because RoB may vary for different outcomes, assessment can be done separately for several outcomes of one study.

The Berger/ISPOR questionnaire contains 33 items grouped into two sections on “relevance” (i.e. applicability) and “credibility” (i.e. internal validity). All key domains of bias are covered, but the questionnaire contains several items on the quality of reporting, such as reporting both absolute and relative effect measures. Therefore, this RoB tool is not recommended for general use.

The Cowley checklist was developed in 1995 as an ad-hoc instrument to be used in a systematic review on total hip replacement. The checklist was created to assess NRS, but the author also prepared similar checklists to assess RCTs and case series. Some of the items in the NRS checklist are disease-specific, which prevents general use of this instrument without prior modification.

The Downs-Black checklist was developed in 1998 (41) and has received widespread international recognition. The instrument can be used to assess the validity of both RCTs and NRS. It consists of 27 items, of which the first 10 items address reporting quality. Questions 11 to 13 relate to external validity. In 2003, Deeks et al. suggested that an item on baseline comparability might be added to the checklist (31). Therefore, this RoB tool obviously requires modification (with refocussing on internal validity only) and cannot be generally recommended any longer.

The EPHPP (Effective Public Health Practice Project Quality Assessment Tool) is intended for use in public health research. It does not include reporting bias and assesses blinding only in general (Data collection valid and reliable?). The items on selection bias are formulated in way that is focused on assessment of external rather than internal validity. Furthermore, the tool contains an item on quality of reporting (Number of withdrawals and drop-outs reported?). Therefore, this RoB tool is not recommended here.

The GRACE checklist consists of 11 items - 6 on data and 5 on methods. Although the checklist is intended primarily for studies on drug therapy and some items point in this direction (“washout period”), the checklist could be used for all types of interventions. One item asks whether study authors were able to “justify the use of a historical comparison group” because for example “it was impossible to identify current users of older treatments”. The fact, however, that a specific study design was impossible does not reduce RoB of a given study. In addition, the domains of detection bias and attrition bias are covered only by one joint item. Therefore, this RoB tool is not recommended here.

MINORS was developed to assess NRS but also RCT evidence. With only 12 items, it is quick to complete. When assessing case series, 4 of these items have to be left blank. However, some items address more than one bias domain. For example, the item on the appropriateness of endpoints aims at detecting both attrition bias and reporting bias. One item of MINORS assesses whether patients in the control group received optimal care. Flaws of examining the wrong research question, however, are not to be considered part of internal validity. Thus, MINORS cannot be generally recommended.

The NOS contains two scales, one on cohort studies and one on case-control studies. In spite of critique from methodologists (49, 50), it has been widely applied in all fields of medicine, because only 8 items are to be scored for cohort studies. The first item of the NOS is about representativeness, thus mixing up internal and external validity. In addition, the NOS lacks an item on reporting bias. Therefore, the NOS appears not recommendable.
With 57 items, the Reisch-Tyson checklist is the longest instrument evaluated here. A large number of items are devoted to the assessment of reporting quality. In the end, a summary score is calculated thus mixing up internal validity and reporting quality. Thus, this instrument, developed already in 1989, cannot be recommended for future use.

Being published in 2013, RoBANS is a quite new instrument. It can be applied to cohort and case-control studies, but not to non-comparative studies. In the journal article by Kim et al. (34), a six-item version of RoBANS is presented, but in 2013 the authors’ institution, the Korean Health Insurance Review & Assessment Service, distributed a brochure which contains RoBANS version 2.0, now containing eight rather than six items. In the new version, the selection of participants is assessed separately for comparisons groups. The second new item deals with “confirmation bias due to inappropriate outcome assessment methods”. Nevertheless, both versions of RoBANS cover all six bias domains, and thus appear recommendable. It should be noted that RoBANS is not intended to produce an overall rating.

The TFCPS instrument is suitable both for RCT and NRS assessment. After completing 26 items in a data collection form, the user has to assess RoB by answering another 23 questions. Some of the questions (e.g. “Was the study population well described?”) address the quality of reporting rather than RoB. Bias due to confounding is covered only partly, because only the appropriateness of statistical analysis is assessed. Reporting bias is not included. Thus, this instrument does not appear recommendable.

<table>
<thead>
<tr>
<th></th>
<th>Confounding bias</th>
<th>Selection bias</th>
<th>Performance bias</th>
<th>Detection bias</th>
<th>Attrition bias</th>
<th>Reporting bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACROBAT</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Berger/ISPOR</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Cowley</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>●</td>
<td>●</td>
<td>-</td>
</tr>
<tr>
<td>Downs-Black</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>EPHPP</td>
<td>●</td>
<td>○</td>
<td>●</td>
<td>○</td>
<td>●</td>
<td>-</td>
</tr>
<tr>
<td>GRACE</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>●</td>
<td>○</td>
<td>-</td>
</tr>
<tr>
<td>MINORS</td>
<td>●</td>
<td>●</td>
<td>-</td>
<td>●</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td>NOS</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>-</td>
</tr>
<tr>
<td>Reisch-Tyson</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>RoBANS</td>
<td>●</td>
<td>○</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>TFCPS</td>
<td>●</td>
<td>○</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 1: RoB checklists’ coverage of key bias domains.
The table contains 6 columns, because one of the 5 bias domains is often split up into two subdomains (bias due to confounding and selection bias).

(● = fully covered; ○ = partly covered; - = not covered; suitable checklists are highlighted by bold print)
2.4. Reliability and ease-of-use of RoB tools

As described in section 2.3, only two checklists appear suitable for use in HTA: ACROBAT-NRSI and RoBANS. Since both instruments are quite new, reliability data are sparse or absent. Testing version 1.0 of RoBANS (containing 6 items) showed moderate reliability (34). No data are available so far on RoBANS version 2.0 and on ACROBAT-NRSI. Nevertheless, detailed instructions are available for both instruments. It is generally recommended to assess RoB in duplicate (i.e. performed by two reviewers independently).

Previous studies on other RoB tools underscore the importance of training in order to achieve adequate reliability in assessments (41, 31, 38, 28). In addition to general training, it may be necessary to define specific aspects of RoB assessments when starting on a new project. HTA researchers should for example agree on the percentage of follow-up completeness which is judged to complete "reasonably complete". Similarly, agreement is required as to which confounders are important to control for. These ad-hoc 'rules' should be reported so that readers can understand RoB assessments in greater detail.

Due to the novelty of both RoB tools, ease-of-use has not been reported yet in the literature. By participating in the pilot testing of ACROBAT-NRSI, the authors of the present guideline gained some practical experience, which showed that ACROBAT-NRSI is relatively easy to use. Endpoint-specific assessments, which are important mainly if blinding or data completeness differs between endpoints, are possible in both RoB tools. In ACROBAT-NRSI, endpoint-specific assessments are standard, while RoBANS offers this feature only as an option. One key advantage of ACROBAT-NRSI is the availability of detailed methodological guidance (http://www.riskofbias.info). In addition, template documents are available to make documentation of the assessment results easier.

2.5. Summarizing the results of RoB assessments

Classification of NRS evidence according to study type is required when assessing RoB using either ACROBAT-NRSI or RoBANS. This is because both instruments contain some items or criteria for item scoring that are specific for different study designs. For ACROBAT-NRSI it is necessary to differentiate between comparative cohort and case-control studies. In addition to these two designs, RoBANS requires that users also recognize cross-sectional and before-after studies. In journal articles, authors of NRS tend to mislabel their studies with regard to study design (51-54), which complicates RoB assessment. As correct identification of study design builds the basis for RoB assessment, this step should already be performed by qualified individuals. Notably, a specific NRS design does not imply higher or lower internal validity as compared to other NRS designs. This represents a shift in methodological thinking for those who have stuck to an overly strict application of evidence levels. Nevertheless, even if cohort studies do not generally have higher RoB than case-control studies and prospective studies are not always ‘better’ than retrospective studies, it still may be appropriate to use design features in literature searches in order to conduct them efficiently.

In instances where both RCT and NRS evidence for the same question is included, researchers should clearly define whether and how NRS can influence the overall internal validity of results. In the Cochrane ACROBAT-NRSI tool, very well performed non-randomised studies can theoretically be judged to have low RoB, which implies that “the study is comparable to a well-performed randomized trial”. However, it remains very doubtful whether this summary rating of "low RoB" can truly be reached when assessing NRS. The authors of ACROBAT-NRSI “anticipate that most NRS will be judged as at least at moderate overall risk of bias.”
Besides internal validity, studies may present other strengths and weaknesses. The GRADE working group has characterized 7 additional domains (beside internal validity) that should be considered when deciding on the overall quality of the evidence (55-57): The certainty of results can be upgraded because of

- a large magnitude of the effect,
- a dose-response gradient, or,
- if residual confounding would have reduced the effect.

Downgrading the certainty of results may be required because of

- inconsistency of results,
- indirectness of evidence,
- imprecision of results, or
- publication bias.

These 7 domains can best be judged at the level of the complete evidence (as summarized over all available studies on an outcome). As these domains can modify the certainty of results, they have an influence on the possible intersection between RCT and NRS evidence. This again suggests that HTA authors when starting a project should think carefully whether the additional domains (especially the possibility of large or very large effects) can counteract the higher RoB of NRS, thereby achieving acceptable certainty of results.

It should be noted that one of the additional domains, indirectness of evidence, is essentially identical to external validity (or applicability). With a focus on RCTs, this issue is addressed in another EunetHTA guideline (35). If the existing RCTs all lack external validity, NRS can sometimes offer a more realistic estimate of the ‘true’ effect, but this estimate does still not reach the overall credibility of high-quality RCT evidence (19, 58-60). Or put simply, low external validity of RCT evidence does not increase the internal validity of NRS evidence. In this context, it is worthwhile mentioning that registry analyses based on so-called ‘real-world data’ can be considered as just one type of cohort study. As such, RoB of registry analyses should be assessed by the same methods as done for any other NRS. Registry analyses come with the promise of minimal selection bias, minimal attrition bias, and a good ability to control for confounding, but their true internal validity may well be lower than that of conventional cohort studies (61).

2.6. Presenting the results of RoB assessments

Presentation of assessment results should be detailed enough to allow readers replication of assessments. As a minimum, domain-specific results should be reported for each study (and each outcome if relevant differences are present). ACROBAT-NRSI offers text fields and requires that assessors enter study-specific explanations and quotations to support their judgment on each bias domain. Ideally, HTA reports include all single items and explanations – preferentially as an appendix. Example tables on how to report domain-specific results can be found at the website of ACROBAT-NRSI. If information is to be presented in a more compact format, summary tables such as the example in Annexe 3 can be prepared.
2.7. Summary of the results

Many of the existing tools for RoB assessment fail to address all relevant domains or mix up assessment of internal validity with quality of reporting. Only two tools, ACROBAT-NRSI and RoBANS, fulfilled the pre-specified criteria, even though reliability of assessments has not yet been demonstrated for these tools. Both tools are suitable for non-randomised comparative studies but not case series. Both tools require at least a basic classification of study design (e.g. cohort study, case-control study). ACROBAT-NRSI appears to have some advantages, such as endpoint-specific assessments, requirement for a summary rating, detailed instructions and documentation guides, and therefore is to be preferred.

Any RoB assessment requires adequate training, especially when assessing NRS. Still, RoB assessment is to some extent an inevitably subjective process. Therefore, detailed reporting of assessment results is necessary in order to make judgmental decisions transparent.
3. Conclusion and main recommendations

The choice between RCT and NRS evidence is loaded with difficult questions. Including non-randomised studies (NRS) in an HTA report clearly requires more resources. Therefore, it is recommended to consider topic-specific circumstances, regulatory context, resources, and timing issues, before making a decision.

If NRS are included, RoB assessment should cover at least 5 different types of bias: selection bias (including bias due to confounding), performance bias, detection bias, attrition bias, and reporting bias. It is important not to mix up internal and external validity. Furthermore, quality of reporting should be kept separately.

Currently, 2 tools are considered the most suitable for RoB assessment of NRS: ACROBAT-NRSI (A Cochrane Risk of Bias Assessment Tool) and RoBANS (Risk of Bias Assessment Tool for Non-randomized Studies). As ACROBAT-NRSI requires endpoint-specific assessments and a summary rating and also offers more detailed instructions and documentation guides as compared to RoBANS, the former tool appears better suited. Therefore, ACROBAT-NRSI is recommended as the currently best tool for assessing RoB of NRS on interventions.

Adequate methodological and clinical knowledge is required for valid and reliable RoB assessment in NRS, because a full understanding of both bias mechanisms and possible confounders is necessary. In addition, clear and consistent decision rules should be agreed on to achieve acceptable reproducibility. Finally, detailed and transparent reporting of assessment methods and results is necessary.
Annexe 1. Bibliography


23. Anglemyer A, Horvath HT, Bero L. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. Cochrane Database Syst Rev 2014;4:Mr000034.


54. Mayo NE, Goldberg MS. When is a case-control study not a case-control study? J Rehabil Med 2009;41:209-16.


Annexe 2. Documentation of literature search

Search engines and sources of information

Search in bibliographic database: Medline (Ovid)

Search date: 06.01.2014

Strategies of research

Ovid MEDLINE(R) In-Process & Other Non-Indexed Citations, January 03, 2014

Ovid MEDLINE(R) 1946 to November Week 3 2013

Ovid MEDLINE(R) Daily Update November 20, 2013

<table>
<thead>
<tr>
<th>#</th>
<th>Searches</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(appraisal* or assessment* or assessing* or rating* or rater*).ti,ab.</td>
</tr>
<tr>
<td>2</td>
<td>(tool* or scale* or checklist*).ti,ab.</td>
</tr>
<tr>
<td>3</td>
<td>((classification* or scoring*) adj1 system*).ti,ab.</td>
</tr>
<tr>
<td>4</td>
<td>Research Design/</td>
</tr>
<tr>
<td>5</td>
<td>st.fs.</td>
</tr>
<tr>
<td>6</td>
<td>4 and 5</td>
</tr>
<tr>
<td>7</td>
<td>or/2-3,6</td>
</tr>
<tr>
<td>8</td>
<td>&quot;Bias (Epidemiology)&quot;/</td>
</tr>
<tr>
<td>9</td>
<td>Reproducibility of Results/</td>
</tr>
<tr>
<td>10</td>
<td>Observer Variation/</td>
</tr>
<tr>
<td>11</td>
<td>((interrater* or interobserver* or observer*) adj3 agreement*).ti,ab.</td>
</tr>
<tr>
<td>12</td>
<td>(bias* or reliability or validity).ti,ab.</td>
</tr>
<tr>
<td>13</td>
<td>quality.ti.</td>
</tr>
<tr>
<td>14</td>
<td>or/8-13</td>
</tr>
<tr>
<td>15</td>
<td>Review Literature as Topic/</td>
</tr>
<tr>
<td>16</td>
<td>Evidence-Based Medicine/</td>
</tr>
<tr>
<td>17</td>
<td>Clinical Trials as Topic/</td>
</tr>
<tr>
<td>18</td>
<td>systematic review*.ti,ab.</td>
</tr>
</tbody>
</table>
Inclusion and exclusion criteria

Inclusion criteria

| I1   | Article contains description or evaluation of a qualitative or quantitative method (e.g. tools, checklists, hierarchies, etc.) to assess risk-of-bias (RoB) |
| I2   | RoB assessment method is intended for at least some type of non-randomised intervention studies |
| I3   | Article type is systematic review, narrative review, or any empirical study (e.g. on validity or reliability of assessment methods) |

Exclusion criteria

| E1   | Article contains only an application of existing RoB assessment methods without formal evaluation of assessment methods |
| E2   | Focus of article is on whether inclusion of non-randomised intervention studies in systematic review is altogether worthwhile |
| E3   | RoB assessment method is intended for diagnostic accuracy studies |
| E4   | Article type is educational review, editorial or letter |
Study selection

Screening of title and abstract was done independently by two researchers. Disagreement was resolved by consensus. Full text document screening was also done independently by two researchers. Disagreement was resolved by consensus or by a third reviewer.

As shown in the PRISMA flow diagram, 33 articles were included. The references of the 38 excluded articles can be found below.

Detailed assessment of included articles

A total of 33 articles were assessed in detail. The 7 studies testing or evaluating existing RoB tools (1-7) were used to support a decision for or against one or another RoB tool. The 3 systematic reviews of existing RoB tools (8-10) were used to identify additional RoB tools.

The remaining 23 articles reported on 20 new (or modified) RoB tools. As described in the following table, all tools (except one) had features that rendered them unsuitable for broader use. Thus, only the tool by Kim et al., RoBANS (Risk of Bias Assessment Tool for Non-randomized Studies), was selected from the bibliographic literature search.
<table>
<thead>
<tr>
<th>Authors, year</th>
<th>Name of tool</th>
<th>Main problem with tool</th>
</tr>
</thead>
<tbody>
<tr>
<td>Berra et al., 2008 (11)</td>
<td>-</td>
<td>Contains several items on external validity and quality of reporting</td>
</tr>
<tr>
<td>Bornhöft et al., (12)</td>
<td>-</td>
<td>Intended only for assessment of external validity</td>
</tr>
<tr>
<td>Chou et al., 2005 (13)</td>
<td>-</td>
<td>Intended only for studies on harms</td>
</tr>
<tr>
<td>Crowe et al., 2011, 2012 (14-17)</td>
<td>CCAT (Crowe Critical Appraisal Tool)</td>
<td>Contains several items on quality of reporting</td>
</tr>
<tr>
<td>Dawson et al., 2013 (18)</td>
<td>Qu-ATEBS (Quality Assessment Tool for Experimental Bruxism Studies)</td>
<td>Ad-hoc tool for use in orthodontics</td>
</tr>
<tr>
<td>Keus et al., 2010 (19)</td>
<td>‘Error matrix approach’</td>
<td>Contains external validity; modification of hierarchy of evidence</td>
</tr>
<tr>
<td>Kim et al., 2013 (20)</td>
<td>RoBANS (Risk of Bias Assessment Tool for Non-randomized Studies)</td>
<td>-</td>
</tr>
<tr>
<td>Kreif et al., 2013 (21)</td>
<td>-</td>
<td>Intended only to assess the confounding domain</td>
</tr>
<tr>
<td>Hrabok et al., 2013 (22)</td>
<td>EBNP (evidence-based neuropsychology) checklist</td>
<td>Ad-hoc tool for use in neuropsychology</td>
</tr>
<tr>
<td>Huisstede et al., 2006 (23)</td>
<td>-</td>
<td>Ad-hoc tool for use in orthopaedic surgery</td>
</tr>
<tr>
<td>Pace et al., 2012 (24)</td>
<td>MMAT (Mixed Methods Appraisal Tool)</td>
<td>Modular design with only 4 items on NRS</td>
</tr>
<tr>
<td>Revuz et al., 2008 (25)</td>
<td>-</td>
<td>Ad-hoc tool for use in dermatology</td>
</tr>
<tr>
<td>Romeiser Logan et al., 2008 (26)</td>
<td>-</td>
<td>Intended only for single case experiments (n-of-1 trials)</td>
</tr>
<tr>
<td>Ross et al., 2011 (27)</td>
<td>SAQOR (Systematic Assessment of Quality in Observational Research)</td>
<td>Contains several items on quality of reporting</td>
</tr>
<tr>
<td>Sirriyeh et al., 2011 (28)</td>
<td>QATSDD (Quality Assessment Tool for Studies with Diverse Designs)</td>
<td>Confounding domain not included; calculation of summary score</td>
</tr>
<tr>
<td>Tate et al., 2008 (29)</td>
<td>SCED (Single-Case Experimental Design)</td>
<td>Intended only for single case experiments (n-of-1 trials)</td>
</tr>
<tr>
<td>Thompson et al., 2011 (30)</td>
<td>-</td>
<td>Only modification of the Downs-and-Black quality checklist</td>
</tr>
<tr>
<td>Tooth et al., 2005 (31)</td>
<td>-</td>
<td>Contains several items on quality of reporting</td>
</tr>
</tbody>
</table>
Tseng et al., 2008 (32) | - | Contains several items on quality of reporting

Wong et al., 2008 (33) | QATSO (Quality Assessment Tool for Systematic Reviews of Observational Studies) | Contains several items on external validity and quality of reporting

References included


References excluded (with reason)

Not I1: Article fails to contain a description or evaluation of a RoB tool (n= 13)


Not I2: Article deals with RoB but not with regard to non-randomised intervention studies


E1: Article contains only an application of existing RoB assessment methods without formal evaluation of assessment methods (n= 3)


E4: Article type is educational review, editorial or letter (n= 13)


12. Yue LQ. Statistical and regulatory issues with the application of propensity score analysis to nonrandomized medical device clinical studies. J Biopharm Stat 2007;17(1):1-13; discussion 5-7, 9-21, 3-7 passim.

### Annexe 3. Example table on how to present RoB assessment results

Table: Outcome-specific risk of bias of non-randomised studies comparing [intervention A] versus [intervention B] and reporting results on [outcome]

<table>
<thead>
<tr>
<th>Study</th>
<th>Bias due to confounding</th>
<th>Bias in selection of participants into the study</th>
<th>Bias in measurement of interventions</th>
<th>Bias due to departure from intended interventions</th>
<th>Bias due to missing data</th>
<th>Bias in measurement of outcomes</th>
<th>Bias in selection of the reported result</th>
<th>Overall bias</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smith et al., 2000</td>
<td>[Low / Moderate / Serious / Critical / NI]</td>
<td>[Low / Moderate / Serious / Critical / NI]</td>
<td>[Low / Moderate / Serious / Critical / NI]</td>
<td>[Low / Moderate / Serious / Critical / NI]</td>
<td>[Low / Moderate / Serious / Critical / NI]</td>
<td>[Low / Moderate / Serious / Critical / NI]</td>
<td>[Low / Moderate / Serious / Critical / NI]</td>
<td>[Low / Moderate / Serious / Critical / NI]</td>
<td>…</td>
</tr>
</tbody>
</table>

NI = No Information

For each individual study reporting data on the outcome of interest, results for each of the 7 bias domains together with the overall RoB should be given. Detailed evaluation forms should be made available in addition (see http://www.riskofbias.info for templates).

Please note that RoB assessment results can best be displayed separately for each outcome. If it is expected that domain-specific RoB results are the same for two or more outcomes, these results might be summarized in one joint table.