eunethta

EUROPEAN NETWORK FOR HEALTH TECHNOLOGY ASSESSMENT

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

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	The recommendation is based on arguments presented in the following parts of the guideline text
1st recommendation: 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.	1.2
2nd recommendation: Internal validity (or risk of bias) should be assessed separately from quality of reporting and external validity (or applicability).	2.1
3rd recommendation: 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.	2.1

4th recommendation: At present, ACROBAT-NRSI (A Cochrane Risk of Bias Assessment Tool) should be used for the RoB assessment of NRS.	2.3
5th recommendation: 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.	2.4
6th recommendation: 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.	2.5
7th recommendation: 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.	2.6

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 founded 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 (NRSMG) 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.

	Confounding	Selection bias	Performance bias	Detection bias	Attrition bias	Reporting bias
ACROBAT	●	●	●	●	●	●
Berger/ISPOR	●	●	●	●	●	●
Cowley	●	●	○	●	●	-
Downs-Black	●	●	●	●	●	●
EPHPP	●	○	●	○	●	-
GRACE	●	●	○	●	○	-
MINORS	●	●	-	●	●	○
NOS	●	●	●	●	●	-
Reisch-Tyson	●	●	●	●	●	●
RoBANS	●	●	●	●	●	●
TFCPS	●	○	●	●	●	-

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

1. Schünemann HJ, Tugwell P, Reeves BC, Akl EA, Santesso N, Spencer FA, et al. Non-randomized studies as a source of complementary, sequential or replacement evidence for randomized controlled trials in systematic reviews on the effects of interventions. Res Synth Methods 2013;4:49-62.
2. Sacks H, Chalmers TC, Smith H, Jr. Randomized versus historical controls for clinical trials. Am J Med 1982;72:233-40.
3. Feinstein AR. An additional basic science for clinical medicine: II. The limitations of randomized trials. Ann Intern Med 1983;99:544-50.
4. Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. BMJ 1998;317:1185-90.
5. Abel U, Koch A. The role of randomization in clinical studies: myths and beliefs. J Clin Epidemiol 1999;52:487-97.
6. McKee M, Britton A, Black N, McPherson K, Sanderson C, Bain C. Methods in health services research. Interpreting the evidence: choosing between randomised and non-randomised studies. BMJ 1999;319:312-5.
7. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. N Engl J Med 2000;342:1878-86.
8. Pocock SJ, Elbourne DR. Randomized trials or observational tribulations? N Engl J Med 2000;342:1907-9.
9. Ioannidis JP, Haidich AB, Lau J. Any casualties in the clash of randomised and observational evidence? BMJ 2001;322:879-80.
10. Ioannidis JP, Haidich AB, Pappa M, Pantazis N, Kokori SI, Tektonidou MG, et al. Comparison of evidence of treatment effects in randomized and nonrandomized studies. JAMA 2001;286:821-30.
11. Lawlor DA, Davey Smith G, Kundu D, Bruckdorfer KR, Ebrahim S. Those confounded vitamins: what can we learn from the differences between observational versus randomised trial evidence? Lancet 2004;363:1724-7.
12. Glasziou P, Chalmers I, Rawlins M, McCulloch P. When are randomised trials unnecessary? Picking signal from noise. BMJ 2007;334:349-51.
13. Brown ML, Gersh BJ, Holmes DR, Bailey KR, Sundt TM, 3rd. From randomized trials to registry studies: translating data into clinical information. Nat Clin Pract Cardiovasc Med 2008;5:613-20.
14. Furlan AD, Tomlinson G, Jadad AA, Bombardier C. Methodological quality and homogeneity influenced agreement between randomized trials and nonrandomized studies of the same intervention for back pain. J Clin Epidemiol 2008;61:209-31.
15. Vandenbroucke JP. Observational research, randomised trials, and two views of medical science. PLoS Med 2008;5:e67.
16. Hoppe DJ, Schemitsch EH, Morshed S, Tornetta P, 3rd, Bhandari M. Hierarchy of evidence: where observational studies fit in and why we need them. J Bone Joint Surg Am 2009;91 Suppl 3:2-9.
17. Concato J, Lawler EV, Lew RA, Gaziano JM, Aslan M, Huang GD. Observational methods in comparative effectiveness research. Am J Med 2010;123:e16-23.

18. Dreyer NA, Tunis SR, Berger M, Ollendorf D, Mattox P, Gliklich R. Why observational studies should be among the tools used in comparative effectiveness research. Health Aff (Millwood) 2010;29:1818-25.

19. van Walraven C, Austin P. Administrative database research has unique characteristics that can risk biased results. J Clin Epidemiol 2012;65:126-31.

20. Etzioni R, Gulati R, Cooperberg MR, Penson DM, Weiss NS, Thompson IM. Limitations of basing screening policies on screening trials: The US Preventive Services Task Force and Prostate Cancer Screening. Med Care 2013;51:295-300.

21. Lauer MS, D'Agostino RB, Sr. The randomized registry trial--the next disruptive technology in clinical research? N Engl J Med 2013;369:1579-81.

22. Melnikow J, LeFevre M, Wilt TJ, Moyer VA. Counterpoint: Randomized trials provide the strongest evidence for clinical guidelines: The US Preventive Services Task Force and Prostate Cancer Screening. Med Care 2013;51:301-3.

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.

24. Dahabreh IJ, Kent DM. Can the learning health care system be educated with observational data? JAMA 2014;312:129-30.

25. Seida J, Dryden DM, Hartling L. The value of including observational studies in systematic reviews was unclear: a descriptive study. J Clin Epidemiol 2014:[Epub ahead of print].

26. Viswanathan M, Ansari M, Berkman N, Chang S, Hartling L, McPheeters L, et al. Assessing the risk of bias of individual studies when comparing medical interventions (AHRQ Methods Guide for Comparative Effectiveness Reviews)2012; AHRQ Publication No. 12-EHC047-EF. Available from: http://effectivehealthcare.ahrq.gov/ehc/products/322/998/MethodsGuideforCERs_Viswanathan_IndividualStudies.pdf.

27. Reeves BC, Higgins JPT, Ramsay C, Shea B, Tugwell P, Wells GA. An introduction to methodological issues when including non-randomised studies in systematic reviews on the effects of interventions. Res Synth Methods 2013;4:1-11.

28. Robertson C, Ramsay C, Gurung T, Mowatt G, Pickard R, Sharma P, et al. Practicalities of using a modified version of the Cochrane Collaboration risk of bias tool for randomised and non-randomised study designs applied in a health technology assessment setting Res Synth Methods 2014;5:200-11.

29. Hartling L, Bond K, Santaguida PL, Viswanathan M, Dryden DM. Testing a tool for the classification of study designs in systematic reviews of interventions and exposures showed moderate reliability and low accuracy. J Clin Epidemiol 2011;64:861-71.

30. Higgins JP, Ramsay C, Reeves BC, Deeks JJ, Shea B, Valentine JC, et al. Issues relating to study design and risk of bias when including non-randomized studies in systematic reviews on the effects of interventions. Res Synth Methods 2013;4:12-25.

31. Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakarovitch C, Song F, et al. Evaluating non-randomised intervention studies. Health Technol Assess 2003;7:iii-x, 1-173.

32. Viswanathan M, Berkman N. Development of the RTI item bank on risk of bias and precision of observational studies (AHRQ Methods Research Report)2011; AHRQ Publication No. 11-EHC028-EF. Available from: http://effectivehealthcare.ahrq.gov/ehc/products/350/784/RTI-Risk-of-Bias_Final-Report_20110916.pdf.

33. Hartling L, Hamm M, Milne A, Vandermeer B, Santaguida PL, Ansari M, et al. Validity and inter-rater reliability testing of quality assessment instruments (AHRQ Methods Research Report)2012 Mar; AHRQ Publication No. 12-EHC039-EF. Available from: http://effectivehealthcare.ahrq.gov/ehc/products/332/1014/Methods-Validity_FinalReport_20120320.pdf.

34. Kim SY, Park JE, Lee YJ, Seo H-J, Sheen S-S, Hahn S, et al. Testing a tool for assessing the risk of bias for nonrandomized studies showed moderate reliability and promising validity. J Clin Epidemiol 2013;66:408-14.

35. EUnetHTA (European Network for Health Technology Assessment). Applicability of evidence in the context of a relative effectiveness assessment of pharmaceuticals.2013. Available from: http://www.eunethta.eu/sites/5026.fedimbo.belgium.be/files/Applicability.pdf.

36. EUnetHTA (European Network for Health Technology Assessment). Internal validity of randomized controlled trials.2013. Available from: http://www.eunethta.eu/sites/5026.fedimbo.belgium.be/files/Internal_Validity.pdf.

37. Ip S, Paulus JK, Balk EM, Dahabreh IJ, Avendano EE, Lau J. Role of single group studies in Agency for Healthcare Research and Quality Comparative Effectiveness Reviews (AHRQ Research White Paper)2013; AHRQ Publication No. 13-EHC007-EF. Available from: http://effectivehealthcare.ahrq.gov/ehc/products/501/1389/White-Paper-Role-of-single-group-studies-1-23-13.pdf.

38. Sanderson S, Tatt ID, Higgins JP. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. Int J Epidemiol 2007;36:666-76.

39. Reisch JS, Tyson JE, Mize SG. Aid to the evaluation of therapeutic studies. Pediatrics 1989;84:815-27.

40. Cowley DE. Prostheses for primary total hip replacement. A critical appraisal of the literature. Int J Technol Assess Health Care 1995;11:770-8.

41. Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. J Epidemiol Community Health 1998;52:377-84.

42. Zaza S, Wright-De Aguero LK, Briss PA, Truman BI, Hopkins DP, Hennessy MH, et al. Data collection instrument and procedure for systematic reviews in the Guide to Community Preventive Services. Task Force on Community Preventive Services. Am J Prev Med 2000;18:44-74.

43. Effective Public Health Practice Project. EPHPP Quality Assessment Tool for Quantitative Studies 2008 [cited 2014]. Available from: http://www.ephpp.ca/tools.html.

44. Wells GA, Shea B, O'Connell D, Peterson J, Welch V, Losos M, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses 2014 [cited 2014]. Available from: http://www.ohri.ca/programs/clinical_epidemiology/oxford.htm.

45. Slim K, Nini E, Forestier D, Kwiatkowski F, Panis Y, Chipponi J. Methodological index for non-randomized studies (MINORS): development and validation of a new instrument. ANZ J Surg 2003;73:712-6.

46. Berger ML, Martin BC, Husereau D, Worley K, Allen JD, Yang W, et al. A questionnaire to assess the relevance and credibility of observational studies to inform health

care decision making: an ISPOR-AMCP-NPC Good Practice Task Force report. Value Health 2014;17:143-56.

47. Dreyer NA, Velentgas P, Westrich K, Dubois R. The GRACE checklist for rating the quality of observational studies of comparative effectiveness: a tale of hope and caution. J Manag Care Pharm 2014;20:301-8.
48. Sterne JAC, Higgins JPT, Reeves BC, on behalf of the development group for ACROBAT-NRSI. A Cochrane Risk Of Bias Assessment Tool: for Non-Randomized Studies of Interventions (ACROBAT-NRSI), Version 1.0.0.2014.
49. Stang A. Critical evaluation of the Newcastle-Ottawa scale for the assessment of the quality of nonrandomized studies in meta-analyses. Eur J Epidemiol 2010;25:603-5.
50. Hartling L, Milne A, Hamm MP, Vandermeer B, Ansari M, Tsertsvadze A, et al. Testing the Newcastle Ottawa Scale showed low reliability between individual reviewers. J Clin Epidemiol 2013;66:982-93.
51. Mihailovic A, Bell CM, Urbach DR. Users' guide to the surgical literature. Case-control studies in surgical journals. Can J Surg 2005;48:148-51.
52. Hellems MA, Kramer MS, Hayden GF. Case-control confusion. Ambul Pediatr 2006;6:96-9.
53. Grimes DA. "Case-control" confusion: mislabeled reports in obstetrics and gynecology journals. Obstet Gynecol 2009;114:1284-6.
54. Mayo NE, Goldberg MS. When is a case-control study not a case-control study? J Rehabil Med 2009;41:209-16.
55. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008;336:924-6.
56. Balshem H, Helfand M, Schünemann HJ, Oxman AD, Kunz R, Brozek J, et al. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol 2011;64:401-6.
57. Berkman ND, Lohr KN, Ansari M, McDonagh M, Balk E, Whitlock E, et al. Grading the strength of a body of evidence when assessing health care interventions for the Effective Health Care Program of the Agency for Healthcare Research and Quality: An Update (AHRQ Methods Guide) 2013; AHRQ Publication No. 13(14)-EHC130-EF. Available from: http://www.effectivehealthcare.ahrq.gov/ehc/products/457/1752/methods-guidance-grading-evidence-131118.pdf.
58. Albert RK. "Lies, damned lies ..." and observational studies in comparative effectiveness research. Am J Respir Crit Care Med 2013;187:1173-7.
59. Freemantle N, Marston L, Walters K, Wood J, Reynolds MR, Petersen I. Making inferences on treatment effects from real world data: propensity scores, confounding by indication, and other perils for the unwary in observational research. BMJ 2013;347:f6409.
60. Lefering R. Strategies for comparative analyses of registry data. Injury 2014;45 Suppl 3:S83-8.
61. Gliklich RE, Dreyer NA. Registries for evaluating patient outcomes - a user's guides, 3rd edition (AHRQ report) 2014; AHRQ Publication No. 13(14)-EHC111. Available from: http://effectivehealthcare.ahrq.gov/ehc/products/420/1897/registries-guide-3rd-edition-vol-1-140430.pdf.

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

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

19	(observational* adj3 studies).ti,ab.
20	or/15-19
21	and/1,7,14,20
22	limit 21 to yr="2005 -Current"

Inclusion and exclusion criteria

Inclusion criteria

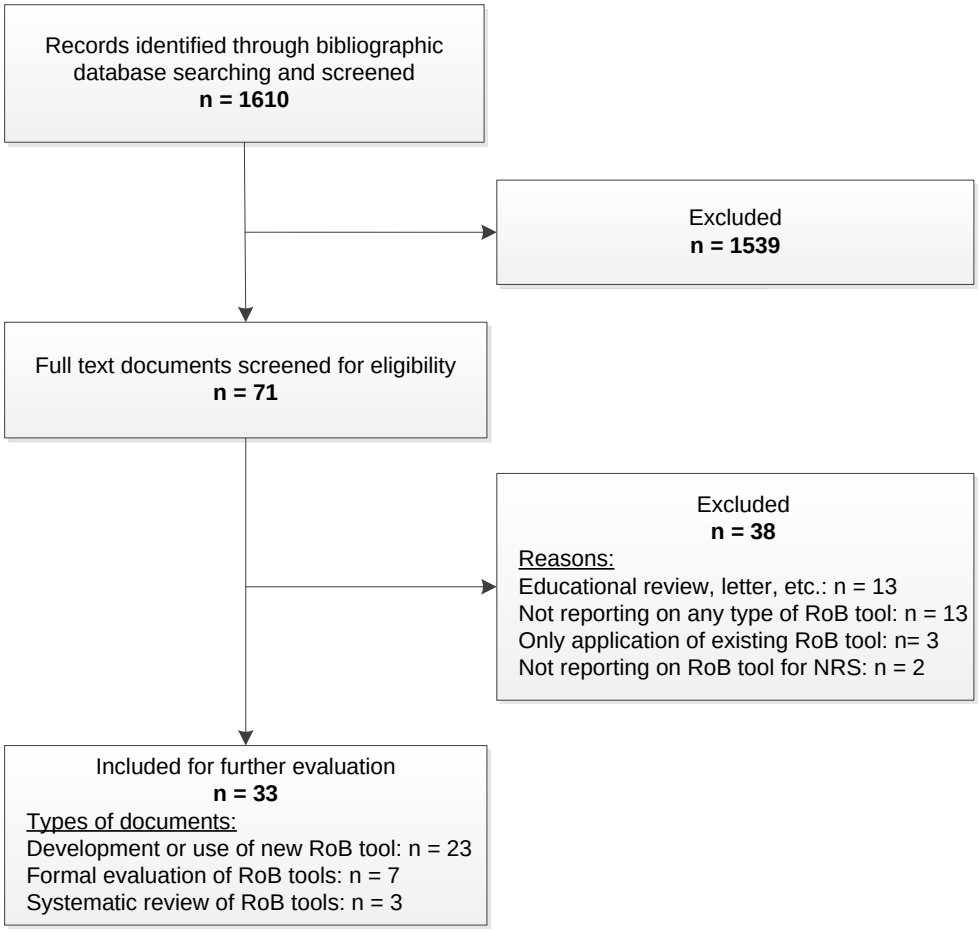
I1	Article contains description or evaluation of a qualitative or quantitative method (e.g. tools, checklists, hierarchies, etc.) to assess risk-of-bias (RoB)
I2	RoB assessment method is intended for at least some type of non-randomised intervention studies
I3	Article type is systematic review, narrative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

Authors, year	Name of tool	Main problem with tool
Berra et al., 2008 (11)	-	Contains several items on external validity and quality of reporting
Bornhöft et al., (12)	-	Intended only for assessment of external validity
Chou et al., 2005 (13)	-	Intended only for studies on harms
Crowe et al., 2011, 2012 (14-17)	CCAT (Crowe Critical Appraisal Tool)	Contains several items on quality of reporting
Dawson et al., 2013 (18)	Qu-ATEBS (Quality Assessment Tool for Experimental Bruxism Studies)	Ad-hoc tool for use in orthodontics
Keus et al., 2010 (19)	‘Error matrix approach’	Contains external validity; modification of hierarchy of evidence
Kim et al., 2013 (20)	RoBANS (Risk of Bias Assessment Tool for Non-randomized Studies)	-
Kreif et al., 2013 (21)	-	Intended only to assess the confounding domain
Hrabok et al., 2013 (22)	EBNP (evidence-based neuropsychology) checklist	Ad-hoc tool for use in neuropsychology
Huisstede et al., 2006 (23)	-	Ad-hoc tool for use in orthopaedic surgery
Pace et al., 2012 (24)	MMAT (Mixed Methods Appraisal Tool)	Modular design with only 4 items on NRS
Revuz et al., 2008 (25)	-	Ad-hoc tool for use in dermatology
Romeiser Logan et al., 2008 (26)	-	Intended only for single case experiments (n-of-1 trials)
Ross et al., 2011 (27)	SAQOR (Systematic Assessment of Quality in Observational Research)	Contains several items on quality of reporting
Sirriyeh et al., 2011 (28)	QATSDD (Quality Assessment Tool for Studies with Diverse Designs)	Confounding domain not included; calculation of summary score
Tate et al., 2008 (29)	SCED (Single-Case Experimental Design)	Intended only for single case experiments (n-of-1 trials)
Thompson et al., 2011 (30)	-	Only modification of the Downs-and-Black quality checklist
Tooth et al., 2005 (31)	-	Contains several items on quality of reporting

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

1. Alperson SY, Berger VW. Opposing systematic reviews: the effects of two quality rating instruments on evidence regarding t'ai chi and bone mineral density in postmenopausal women. J Altern Complement Med 2011;17(5):389-95.
2. Hartling L, Fernandes RM, Seida J, Vandermeer B, Dryden DM. From the trenches: a cross-sectional study applying the GRADE tool in systematic reviews of healthcare interventions. PLoS ONE 2012;7(4):e34697.
3. Hartling L, Milne A, Hamm MP, Vandermeer B, Ansari M, Tsertsvadze A, et al. Testing the Newcastle Ottawa Scale showed low reliability between individual reviewers. J Clin Epidemiol 2013;66(9):982-93.
4. Oremus M, Oremus C, Hall GBC, McKinnon MC, Ect, Cognition Systematic Review T. Inter-rater and test-retest reliability of quality assessments by novice student raters using the Jadad and Newcastle-Ottawa Scales. BMJ Open 2012;2(4).
5. Voss PH, Rehfuess EA. Quality appraisal in systematic reviews of public health interventions: an empirical study on the impact of choice of tool on meta-analysis. J Epidemiol Community Health 2013;67(1):98-104.
6. Wiart L, Kolaski K, Butler C, Vogtle L, Logan LR, Hickman R, et al. Interrater reliability and convergent validity of the American Academy for Cerebral Palsy and Developmental Medicine methodology for conducting systematic reviews. Dev Med Child Neurol 2012;54(7):606-11.
7. Baker A, Young K, Potter J, Madan I. A review of grading systems for evidence-based guidelines produced by medical specialties. Clin Med 2010;10(4):358-63.
8. Dreier M, Borutta B, Stahmeyer J, Krauth C, Walter U. Comparison of tools for assessing the methodological quality of primary and secondary studies in health technology assessment reports in Germany. GMS Health Technol Assess. 2010;6:Doc07.
9. Neyarapally GA, Hammad TA, Pinheiro SP, Iyasu S. Review of quality assessment tools for the evaluation of pharmacoepidemiological safety studies. BMJ Open 2012;2(5).
10. Sanderson S, Tatt ID, Higgins JPT. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. Int J Epidemiol 2007;36(3):666-76.
11. Berra S, Elorza-Ricart JM, Estrada M-D, Sánchez E. [A tool [corrected] for the critical appraisal of epidemiological cross-sectional studies]. Gac Sanit 2008;22(5):492-7.
12. Bornhöft G, Maxion-Bergemann S, Wolf U, Kienle GS, Michalsen A, Vollmar HC, et al. Checklist for the qualitative evaluation of clinical studies with particular focus on external validity and model validity. BMC Med Res Methodol 2006;6:56.

13. Chou R, Helfand M. Challenges in systematic reviews that assess treatment harms. Ann Intern Med 2005;142(12 Pt 2):1090-9.

14. Crowe M, Sheppard L. A general critical appraisal tool: an evaluation of construct validity. Int J Nurs Stud 2011;48(12):1505-16.

15. Crowe M, Sheppard L, Campbell A. Comparison of the effects of using the Crowe Critical Appraisal Tool versus informal appraisal in assessing health research: a randomised trial. Int J Evid Based Healthc 2011;9(4):444-9.

16. Crowe M, Sheppard L, Campbell A. Reliability analysis for a proposed critical appraisal tool demonstrated value for diverse research designs. J Clin Epidemiol 2012;65(4):375-83.

17. Crowe M, Sheppard L. A review of critical appraisal tools show they lack rigor: Alternative toolstructure is proposed. J Clin Epidemiol 2011;64(1):79-89.

18. Dawson A, Raphael KG, Glaros A, Axelsson S, Arima T, Ernberg M, et al. Development of a quality-assessment tool for experimental bruxism studies: reliability and validity. J Orofac Pain 2013;27(2):111-22.

19. Keus F, Wetterslev J, Gluud C, van Laarhoven CJHM. Evidence at a glance: error matrix approach for overviewing available evidence. BMC Med Res Methodol 2010;10:90.

20. Kim SY, Park JE, Lee YJ, Seo H-J, Sheen S-S, Hahn S, et al. Testing a tool for assessing the risk of bias for nonrandomized studies showed moderate reliability and promising validity. J Clin Epidemiol 2013;66(4):408-14.

21. Kreif N, Grieve R, Sadique MZ. Statistical methods for cost-effectiveness analyses that use observational data: a critical appraisal tool and review of current practice. Health Econ 2013;22(4):486-500.

22. Hrabok M, Dykeman J, Sherman EMS, Wiebe S. An evidence-based checklist to assess neuropsychological outcomes of epilepsy surgery: how good is the evidence? Epilepsy Behav 2013;29(3):443-8.

23. Huisstede BMA, Miedema HS, van Opstal T, de Ronde MTM, Kuiper JI, Verhaar JAN, et al. Interventions for treating the posterior interosseus nerve syndrome: a systematic review of observational studies. J Peripher Nerv Syst 2006;11(2):101-10.

24. Pace R, Pluye P, Bartlett G, Macaulay AC, Salsberg J, Jagosh J, et al. Testing the reliability and efficiency of the pilot Mixed Methods Appraisal Tool (MMAT) for systematic mixed studies review. Int J Nurs Stud 2012;49(1):47-53.

25. Revuz J, Moyse D, Poli F, Pawin H, Faure M, Chivot M, et al. A tool to evaluate rapidly the quality of clinical trials on topical acne treatment. J Eur Acad Dermatol Venereol 2008;22(7):800-6.

26. Romeiser Logan L, Hickman RR, Harris SR, Heriza CB. Single-subject research design: recommendations for levels of evidence and quality rating.[Erratum appears in Dev Med Child Neurol 2009 Mar;51(3):247]. Dev Med Child Neurol 2008;50(2):99-103.

27. Ross LE, Grigoriadis S, Mamisashvili L, Koren G, Steiner M, Dennis CL, et al. Quality assessment of observational studies in psychiatry: an example from perinatal psychiatric research. Int J Methods Psychiatr Res 2011;20(4):224-34.

28. Sirriyeh R, Lawton R, Gardner P, Armitage G. Reviewing studies with diverse designs: the development and evaluation of a new tool. J Eval Clin Pract 2012;18(4):746-52.

29. Tate RL, McDonald S, Perdices M, Togher L, Schultz R, Savage S. Rating the methodological quality of single-subject designs and n-of-1 trials: introducing the Single-Case Experimental Design (SCED) Scale. Neuropsychol 2008;18(4):385-401.

30. Thompson S, Ekelund U, Jebb S, Lindroos AK, Mander A, Sharp S, et al. A proposed method of bias adjustment for meta-analyses of published observational studies. Int J Epidemiol 2011;40(3):765-77.

31. Tooth L, Ware R, Bain C, Purdie DM, Dobson A. Quality of reporting of observational longitudinal research. Am J Epidemiol 2005;161(3):280-8.

32. Tseng TY, Breau RH, Fesperman SF, Vieweg J, Dahm P. Evaluating the evidence: the methodological and reporting quality of comparative observational studies of surgical interventions in urological publications. BJU Int 2009;103(8):1026-31.

33. Wong WCW, Cheung CSK, Hart GJ. Development of a quality assessment tool for systematic reviews of observational studies (QATSO) of HIV prevalence in men having sex with men and associated risk behaviours. Emerg Themes Epidemiol 2008;5:23.

References excluded (with reason)

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

1. Baron G, Boutron I, Giraudeau B, Ravaud P. Violation of the intent-to-treat principle and rate of missing data in superiority trials assessing structural outcomes in rheumatic diseases. Arthritis Rheum 2005;52(6):1858-65.

2. Berger ML, Dreyer N, Anderson F, Towse A, Sedrakyan A, Normand SL. Prospective observational studies to assess comparative effectiveness: the ISPOR Good Research Practices Task Force report. Value Health 2012;15(2):217-30.

3. Blackman KC, Zoellner J, Berrey LM, Alexander R, Fanning J, Hill JL, et al. Assessing the internal and external validity of mobile health physical activity promotion interventions: a systematic literature review using the RE-AIM framework. J Med Internet Res 2013;15(10):e224.

4. Bossuyt PMM, Besselink MGH. Is there such a thing as 'fitting evidence'? [Dutch]. Ned Tijdschr Geneeskd 2013;157(15):A6027.

5. Concato J. Study design and "evidence" in patient-oriented research. Am J Respir Crit Care Med 2013;187(11):1167-72.

6. Foster MJ, Shurtz S. Making the Critical Appraisal for Summaries of Evidence (CASE) for evidence-based medicine (EBM): critical appraisal of summaries of evidence. J Med Libr Assoc 2013;101(3):192-8.

7. Jolliffe D, Murray J, Farrington D, Vannick C. Testing the Cambridge Quality Checklists on a review of disrupted families and crime. Crim Behav Ment Health 2012;22(5):303-14.

8. O'Connor DP, Brinker MR. Challenges in outcome measurement: clinical research perspective. Clin Orthop Rel Res 2013;471(11):3496-503.

9. Elissen AMJ, Adams JL, Spreeuwenberg M, Duimel-Peeters IGP, Spreeuwenberg C, Linden A, et al. Advancing current approaches to disease management evaluation: capitalizing on heterogeneity to understand what works and for whom. BMC Med Res Methodol 2013;13:40.

10. Gugiu PC, Gugiu MR. A critical appraisal of standard guidelines for grading levels of evidence. Eval Health Prof 2010;33(3):233-55.
11. Kamioka H, Kawamura Y, Tsutani K, Maeda M, Hayasaka S, Okuizum H, et al. A checklist to assess the quality of reports on spa therapy and balneotherapy trials was developed using the Delphi consensus method: the SPAC checklist. Complement Ther Med 2013;21(4):324-32.
12. Stang A. Critical evaluation of the Newcastle-Ottawa scale for the assessment of the quality of nonrandomized studies in meta-analyses. Eur J Epidemiol 2010;25(9):603-5.
13. Vanderweele TJ, Arah OA. Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. Epidemiology 2011;22(1):42-52.

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

1. Brouwers MC, Johnston ME, Charette ML, Hanna SE, Jadad AR, Browman GP. Evaluating the role of quality assessment of primary studies in systematic reviews of cancer practice guidelines. BMC Med Res Methodol 2005;5(1):8.
2. Dixon E, Hameed M, Sutherland F, Cook DJ, Doig C. Evaluating meta-analyses in the general surgical literature: a critical appraisal. Ann Surg 2005;241(3):450-9.
3. Hirji KF. No short-cut in assessing trial quality: a case study. Trials 2009;10:1.
4. Loudon K, Zwarenstein M, Sullivan F, Donnan P, Treweek S. Making clinical trials more relevant: improving and validating the PRECIS tool for matching trial design decisions to trial purpose. Trials 2013;14:115.
5. Shamliyan T, Kane RL, Dickinson S. A systematic review of tools used to assess the quality of observational studies that examine incidence or prevalence and risk factors for diseases. J Clin Epidemiol 2010;63(10):1061-70.
6. Shamliyan T, Kane RL, Jansen S. Systematic reviews synthesized evidence without consistent quality assessment of primary studies examining epidemiology of chronic diseases. J Clin Epidemiol 2012;65(6):610-8.
7. Terwee CB, Mokkink LB, Knol DL, Ostelo RWJG, Bouter LM, De Vet HCW. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Qual Life Res 2012;21(4):651-7.
8. Yang X, Shoptaw S. Assessing missing data assumptions in longitudinal studies: an example using a smoking cessation trial. Drug Alcohol Depend 2005;77(3):213-25.
9. Furlan AD, Pennick V, Bombardier C, van Tulder M, Cochrane Back Review Group. 2009 updated method guidelines for systematic reviews in the Cochrane Back Review Group. Spine 2009;34(18):1929-41.

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

1. Nye C, Hahs-Vaughn D. Assessing methodological quality of randomized and quasi-experimental trials: a summary of stuttering treatment research. Int J Speech Lang Pathol 2011;13(1):49-60.
2. Hamre HJ, Glockmann A, Kienle GS, Kiene H. Combined bias suppression in single-arm therapy studies. J Eval Clin Pract 2008;14(5):923-9.
3. Pearson M, Peters J. Outcome reporting bias in evaluations of public health interventions: evidence of impact and the potential role of a study register. J Epidemiol Community Health 2012;66(4):286-9.

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

1. Akobeng AK. Assessing the validity of clinical trials. J Ped Gastroent Nutrition 2008;47(3):277-82.
2. Callas PW. Searching the biomedical literature: research study designs and critical appraisal. Clin Lab Sci 2008;21(1):42-8.
3. Guzelian PS, Victoroff MS, Halmes NC, James RC, Guzelian CP. Evidence-based toxicology: a comprehensive framework for causation. Human Exp Toxicol 2005;24(4):161-201.
4. Hiebert R, Nordin M. Methodological aspects of outcomes research. Eur Spine J 2006;15(Suppl 1):S4-S16.
5. Lang S, Kleijnen J. Quality assessment tools for observational studies: lack of consensus. Int J Evid Based Healthcare 2010;8(4):247.
6. Moyer A, Finney JW. Rating methodological quality: toward improved assessment and investigation. Account Res 2005;12(4):299-313.
7. Paradis C. Bias in surgical research. Ann Surg 2008;248(2):180-8.
8. Urschel JD. How to analyze an article. World J Surg 2005;29(5):557-60.
9. Hannan EL. Randomized clinical trials and observational studies: guidelines for assessing respective strengths and limitations. JACC Cardiovasc Interv 2008;1(3):211-7.
10. Levenson MS, Yue LQ. Regulatory issues of propensity score methodology application to drug and device safety studies. J Biopharm Stat 2013;23(1):110-21.
11. Luo Z, Gardiner JC, Bradley CJ. Applying propensity score methods in medical research: pitfalls and prospects. Med Care Res Rev 2010;67(5):528-54.
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.
13. Manchikanti L, Singh V, Smith HS, Hirsch JA. Evidence-based medicine, systematic reviews, and guidelines in interventional pain management: part 4: observational studies. Pain Physician 2009;12(1):73-108.

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]

Study	Bias due to confounding	Bias in selection of participants into the study	Bias in measurement of interventions	Bias due to departures from intended interventions	Bias due to missing data	Bias in measurement of outcomes	Bias in selection of the reported result	Overall bias	Comments
Smith et al., 2000	[Low / Moderate / Serious / Critical / NI]	[Low / Moderate / Serious / Critical / NI]	[Low / Moderate / Serious / Critical / NI]	[Low / Moderate / Serious / Critical / NI]	[Low / Moderate / Serious / Critical / NI]	[Low / Moderate / Serious / Critical / NI]	[Low / Moderate / Serious / Critical / NI]	[Low / Moderate / Serious / Critical / NI]	…
…									

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.