eunethta

EUROPEAN NETWORK FOR HEALTH TECHNOLOGY ASSESSMENT

GUIDELINE

LEVELS OF EVIDENCE

Internal validity of randomized controlled trials

Final version

February 2013

The primary objective of EUnetHTA JA1 WP5 methodology guidelines is to focus on methodological challenges that are encountered by HTA assessors while performing a rapid relative effectiveness assessment of pharmaceuticals.

This guideline “Levels of evidence: Internal validity (of randomized controlled trials)”has been elaborated by experts from CAST/SDU and IQWiG, reviewed and validated by all members of WP5 of the EUnetHTA network; the whole process was coordinated by HAS. 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.

Table of contents

Acronyms – Abbreviations

ADAS-cog - Alzheimer's Disease Assessment Scale Cognitive subscale

AE - Adverse event

AMSTAR - A Measurement Tool to Assess Systematic Reviews

CAST/SDU – Centre for Applied Health Services Research and Technology Assessment, University of Southern Denmark

CI - Confidence interval

CONSORT - Consolidated Standards of Reporting Trials

EUnetHTA - European network for Health Technology Assessment

GRADE - Grading of Recommendations Assessment, Development and Evaluation

HAS - Haute Autorité de Santé

HTA - Health technology assessment

INAHTA - International Network of Agencies for Health Technology Assessment

IQWiG - Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (Institute for Quality and Efficiency in Health Care)

ITT - Intention to treat

OQAQ - Overview Quality Assessment Questionnaire

OR - Odds ratio

PHARMAC - Pharmaceutical Management Agency

PRISMA - Preferred Reporting Items for Systematic Reviews and Meta-Analyses

RCT - Randomized controlled trial

REA - Relative effectiveness assessment

SAE - Serious adverse event

STROBE - Strengthening the Reporting of Observational Studies in Epidemiology

WP5 – Work package 5

Summary and recommendations

Summary

Internal validity describes the extent to which the (treatment) difference observed in a trial (or a meta-analysis) is likely to reflect the 'true' effect within the trial (or in the trial population) by considering methodological quality criteria. Because the 'truth' can never be assessed, it is more appropriate to speak of the potential for or risk of bias.

The present guideline focuses on the assessment of the risk of bias of randomized controlled trials (RCTs), the most relevant trials for relative effectiveness assessment (REA) of pharmaceuticals. The quality assessment of non-randomized and diagnostic accuracy studies will be elaborated in separate guidelines. Likewise, a separate guideline deals with the problem of assessing applicability.

Over the years, the Cochrane Collaboration has developed an elaborate framework to assess the risk of bias in RCTs (Higgins et al. 2011). This framework aims to inform readers of systematic reviews about the trustworthiness of the results. It is based on both theoretical considerations and empirical evidence of 5 major types of bias. It can be regarded as a generally accepted standard, or 'gold standard', and its use has been advocated by a number of HTA agencies active in EUnetHTA. Hence, for the present guideline it is appropriate not to conduct an extensive literature search and to refer mainly to the Cochrane risk of bias tool.

The different types of potential bias can be separated into at least 6 categories: selection, performance, detection, attrition, reporting, and other sources of bias. With regard to these different types of bias and the strategies used in clinical trials to protect from such bias, the Cochrane Handbook for Systematic Reviews of Interventions ('Cochrane Handbook') specifies the following 7 relevant domains for the assessment of the risk of bias: random sequence generation, allocation concealment, blinding of participants and personnel, blinding of outcome assessment, incomplete outcome data, selective reporting, and other sources of bias (Higgins & Green 2011).

The risk of bias should be assessed on 2 levels, i.e. firstly, on a (general) study level, and secondly, on an outcome level. For example, selection and performance bias threaten the validity of the entire study, while the other types of bias may be outcome specific. The risk of bias is then categorized into 3 groups: low risk of bias, high risk of bias, and unclear risk of bias.

There are at least 4 options to deal with the risk of bias: (i) rely only on studies with a low risk of bias; (ii) perform sensitivity analyses according to the different risk of bias categories; (iii) describe the uncertainty with regard to the different levels of risk of bias, so that subsequent decisions can be made considering this uncertainty; (iv) combine option (ii) and (iii).

If an REA is not or not fully based on primary studies, but rather on systematic reviews (e.g. due to limited resources), it is also necessary to assess whether the underlying systematic review(s) has/have only minimal methodological flaws. Various instruments exist to assess the quality of systematic reviews (Shea et al. 2007). However, only a minority of these instruments are formally validated, widely used, and focused on methodological quality rather than on reporting quality. If an REA is to be performed on the basis of systematic reviews rather than on primary studies, it is strongly recommended that the methodological quality of the underlying reviews is assessed, either by the Oxman and Guyatt index (Oxman & Guyatt 1991), or by 'A Measurement Tool to Assess Systematic Reviews' (AMSTAR) (Shea 2007). If the quality does not exceed a pre-specified threshold (e.g. at least 5 of 7 possible points in the overall assessment of the Oxman and Guyatt index), the corresponding systematic review should not be used as a basis for the REA. It is then necessary to conduct a separate systematic review for the underlying research question (with an assessment of the internal validity of the identified primary studies according to this guideline).

Recommendations

Recommendation 1 Use the risk of bias concept of the Cochrane Collaboration to assess the internal validity of RCTs within an REA. Chapter 8 and table 8.5.d of the Cochrane Handbook (Higgins & Green 2011) provide detailed guidance.
Recommendation 2 Provide appropriate training and clear and consistent decision rules to achieve acceptable reproducibility of the risk of bias assessments. The use of standardized extraction sheets is also recommended.
Recommendation 3 Within an REA, specify in advance how to deal with studies with a high or unclear risk of bias. There are at least 4 options: (i) rely only on studies with a low risk of bias; (ii) perform sensitivity analyses according to the different risk of bias categories; (iii) describe the uncertainty with regard to the different levels of risk of bias, so that subsequent decisions can be made considering this uncertainty; (iv) combine option (ii) and (iii).
Recommendation 4 Use a validated tool to assess the methodological quality of systematic reviews: the Oxman and Guyatt index (Oxman & Guyatt 1991, Jadad & Murray 2007) and the AMSTAR instrument (Shea et al. 2007) are recommended. Both instruments are useful, without a preference for either one.

1. Introduction

1.1. Definitions

- **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.
- **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 (Higgins & Green 2011).
- **Relative effectiveness:** can be defined as the extent to which an intervention does more good than harm, compared to one or more intervention alternatives for achieving the desired results, when provided under the usual circumstances of health care practice (Pharmaceutical Forum 2008).
- **Systematic reviews**: publications that summarize and assess the results of primary studies in a systematic, reproducible, and transparent way.
- **Health technology assessment**: a multidisciplinary process that summarizes information about the medical, social, economic and ethical issues related to the use of a health technology in a systematic, transparent, unbiased, robust manner. Its aim is to inform the formulation of safe, effective, health policies that are patient focused and seek to achieve best value (EUnetHTA 2012).
- **(Single) Rapid assessment of relative effectiveness of pharmaceuticals**: defined as rapid assessment of a new technology at the time of introduction to the market and comparing the new technology to standard care. This will be referred to hereafter as the ***Rapid Assessment***.
- **Full assessment of relative effectiveness of pharmaceuticals**: defined as full assessment (non-rapid) of (all) available technologies for a particular step in a treatment pathway for a specific condition. This will be referred to hereafter as the ***Full Assessment***.

1.2. Context

1.2.1. Problem statement

To what extent can it be assessed whether the data from a study (e.g. an RCT) or a collection of studies (e.g. a meta-analysis within an REA) are likely to reflect the 'truth' by considering methodological quality criteria? This is essential to allow conclusions about the certainty (or uncertainty) of results for subsequent support of decision-making processes.

1.2.2. Discussion (on the problem statement)

Internal validity describes the extent to which the (treatment) difference observed in a trial (or a meta-analysis) is likely to reflect the 'true' effect within the trial (or in the trial population) by considering methodological quality criteria. Because the 'truth' can never be assessed, it is more appropriate to speak of the potential for or **risk of bias**. Internal validity has to be differentiated from external validity – or better – applicability, which is the topic of a separate guideline.

Over the years, the ***Cochrane Collaboration*** has developed an elaborate framework to assess the risk of bias in RCTs (Higgins et al. 2011). This framework aims to inform readers of systematic reviews about the trustworthiness of the results. It is based on both theoretical considerations and empirical evidence of the potential impact of the different types of bias. It can be regarded as a generally accepted standard, or 'gold standard', and its use has been advocated by a number of HTA agencies active in EUnetHTA. Hence, for the present guideline it is appropriate not to conduct an extensive literature search and to refer mainly to the Cochrane risk of bias tool.

Another important framework for the assessment of the quality of evidence was developed by the ***GRADE*** (Grading of Recommendations Assessment, Development and Evaluation) working group. This framework combines aspects of both internal and external validity, but also of the precision of estimates, the magnitude of effects, and the consistency of results within one single approach to grade the **'quality of the body of evidence'**. Because the scope of the GRADE approach goes beyond the assessment of the single domain 'internal validity' or 'risk of bias', the present guideline focuses on the Cochrane risk of bias tool. Nevertheless, the concept of risk of bias is incorporated within the GRADE framework, so that there is virtually no difference in assessing 'internal validity' between the 2 approaches.

The current guideline focuses on the assessment of the risk of bias of RCTs, the most relevant trials for REA of pharmaceuticals; non-randomized studies – if used for the evaluation of effects of interventions within the REA – inevitably carry a high risk of selection bias and subsequent confounding. Furthermore, non-randomized studies are mostly unblinded, and the intention-to-treat (ITT) principle is even more difficult to realize. Nevertheless, it is useful to assess the quality of evidence from non-randomized studies if the decision was made to include those studies in an REA, notably a full assessment. The quality assessment of non-randomized studies goes beyond the risk of bias assessment of RCTs, because special attention has to be paid to whether and how possible confounders were dealt with in the absence of randomization (e.g. pre-definition of possible confounders, adjustment procedures, matching, etc.). Moreover, there are many types of non-randomized studies (e.g. [observational] cohort studies, case-control studies, uncontrolled before-after studies, interrupted-time-series studies, and [interventional] controlled trials using other allocation strategies than randomization), which may require different instruments for assessing internal validity. The quality assessment of non-randomized studies will therefore be elaborated in a separate guideline, the scope of which will also cover rapid and full assessment of non-pharmaceutical (interventional) health technologies.

1.3. Scope/Objective(s) of the guideline

The guideline aims to provide recommendations for the assessment of the internal validity of RCTs whose purpose is the determination of the relative effectiveness of pharmaceuticals. It does not aim to provide recommendations for the quality assessment of non-randomized studies or diagnostic accuracy studies. Both issues will be addressed in separate guidelines. Likewise, a separate guideline deals with the problem of assessing applicability. However, some recommendations are given for the case when an REA is not or not fully based on primary studies, but rather on one or more systematic review(s).

1.4. Related EUnetHTA documents

This guideline should be read in conjunction with the following documents:

1. EUnetHTA guideline on levels of evidence: applicability of evidence in the context of a relative effectiveness assessment of pharmaceuticals

2. Analysis and synthesis of literature

2.1. Analysis of the literature

Because the Cochrane risk of bias tool can be regarded as a generally accepted standard, it is largely referred to in the subsequent sections, and an extensive literature search was not conducted.

2.2. Summary of the results

2.2.1. Types of bias

The different types of possible bias can be separated into at least 6 categories:

- selection bias,
- performance bias,
- detection bias,
- attrition bias,
- reporting bias,
- other sources of bias.

Selection bias may arise if patient characteristics are (relevantly) different between the treatment groups to be compared. If such a characteristic is related both to the outcome(s) of interest and the selection of treatment, then it is a confounder. If confounding takes place, group differences with respect to the outcome(s) of interest cannot be definitely separated between an effect generated by the treatment or by confounding. In addition, observed treatment differences may be diminished by confounding.

Selection bias can be minimized if the allocation of the patients to the treatment groups occurs by chance, which will be guaranteed by true **randomization**. Randomization itself has 2 important components: the generation of the random **allocation sequence** and the **concealment** of the allocation before inclusion of patients in a trial. If the allocation sequence is known to the person who decides on the inclusion of patients in the trial before inclusion, selective non-inclusion of patients who in fact fulfil the in- and exclusion criteria may occur. One of the first meta-epidemiological studies investigating the empirical evidence of bias observed clearly exaggerated treatment effect estimates in trials with inappropriate or even unclear concealment in comparison to those with adequate concealment (Schulz et al. 1995). In similar meta-epidemiological studies, however, this exaggeration decreased over time (Herbison et al. 2011). Nevertheless, trials with clearly inadequate concealment (e.g. alternate allocation by day of week or year of birth) are regarded as not truly randomized by many HTA agencies and therefore excluded from the pool of genuine RCTs.

Performance bias may arise if the concomitant care of patients within a study is different between the treatment groups. Possible performance bias can be decreased by keeping the applied treatment of interest blinded during the trial. **Blinding** is possible for different players within a trial: treating physicians, other caregivers, patients, and outcome assessors. If nobody knows the applied treatments, the study is often designated as a double-blind trial.[1] However, it should be noted that there is not a real common understanding of the term ‘double blinding’. In some cases trials are designated as double blind only because 2 parties (e.g. the patients and the outcome assessors) are blinded, while others (e.g. the treating physicians) are in fact not. The term ‘single blind’ is used for studies where subjects, but not investigators or outcome assessors, are blinded.

[1] The term ‘triple-blind’ is sometimes used if it is intended to highlight that the persons who are involved in data management (i.e., data managers and biostatisticians) are also kept blinded.

A trial without any blinding is usually designated as an **open (label) trial**. However, if the ideal of total blinding cannot be achieved (e.g. because of typical side effects), it is often possible to keep single players blinded, e.g. the patients or the outcome assessors.

Detection bias may arise if the outcomes of interest are differently assessed between treatment groups – consciously or subconsciously. Like performance bias, the risk of detection bias can be minimized by **blinding**. Again, if 'double blinding' cannot be achieved, it is usually possible to keep the outcome assessors blinded. The necessity for blinding will mostly depend on the nature of the outcome of interest: while it is mandatory for a proper assessment of so-called subjective endpoints (e.g. patient-reported outcomes such as pain or quality of life) (EMA 2005, FDA 2009), it may be less critical if so-called objective endpoints such as mortality are assessed.

There is definite empirical evidence that unblinded or inadequately blinded trials carry a high risk of bias for subjective outcomes (Wood 2008, Hróbjartsson 2012). It should be noted that many investigator-assessed outcomes also have a subjective component, for example, if interpretation of imaging is essential for determining the outcome. Outcome assessment by independent (and – if possible – blinded) adjudication committees is a helpful design instrument in such situations (e.g. central independent adjudication committee review of radiographic images to determine whether the pre-defined definition of an outcome/adverse event was fulfilled). Furthermore, the occurrence of an event (e.g. progression of a disease) or time to this event is sometimes the outcome of interest or part of the outcome of interest. If some specific investigations at follow-up(s) are necessary to assess this event in such a situation (e.g. assessment of progression-free survival in oncology), it is essential to guarantee some standardization and ensure that the timing of follow-up(s) is equal between the treatment groups in open trials (EMA 2011).

Attrition bias may arise if an important proportion of patients are lost for the statistical analysis due to different reasons, e.g. lost to follow-up, withdrawals, missing values or protocol violations. Such reasons carry a potential bias, because they have the risk of being related to both the characteristics of patients relevant for the outcome of interest and the applied treatment, and therefore may introduce selective attrition in the analysed population.

For example, analyses of outcomes might be considered as invalid if more than 30% of patients are not included, or if the difference in excluded patients exceeds the absolute value of 15%. However, even smaller proportions of excluded patients may lead to serious bias, if the group difference is small and potentially outweighed by the proportions of excluded patients. For binary outcome data it can be generally stated that the importance of particular rates of excluded patients is dependent on the relation between the number of excluded patients and the number of events in the intervention and control groups. Similar considerations can be made for continuous outcomes. This means that the above-mentioned thresholds for an 'acceptable' exclusion rate (30%) or difference in exclusion rates (15%) should be understood as an initial approximation. In certain circumstances deviations above or below these figures may be appropriate.

The most important instrument to deal with possible attrition bias is the **ITT principle**, i.e. the principle of analysing all patients within a trial according to their allocated treatment group. However, this principle is often difficult to apply, because in nearly every trial missing values for the outcomes of interest occur. Therefore, it is sometimes necessary to apply a strategy for the replacement of missing values to enable an ITT analysis. However, such replacement strategies themselves carry a risk of bias. So it is important to apply a replacement strategy that does not lead to anti-conservative treatment effect estimates, i.e. in the direction of the statistical alternative hypothesis.

It should be noted in this respect that 'conservatism' does not mean the same in superiority and non-inferiority (or equivalence) trials: in a superiority trial a replacement strategy which diminishes group differences (e.g. by assigning all lost patients to 'failures' or 'successes') may lead to a conservative estimate (in favour of the null hypothesis), while the same strategy may lead to an anti-conservative estimate (in favour of the alternative hypothesis) in non-inferiority or equivalence trials. Therefore, replacement strategies should be adapted to the underlying research hypothesis (Lange 2001).

However, the corresponding 'behaviour' of replacement strategies depends also on drop-out mechanisms and the (natural) course of the disease (Unnebrink & Windeler) as well as on the

influence of the strategy on variance estimates by increasing or decreasing the variance. **Sensitivity analyses** are in general useful in assessing whether the results are robust if different replacement strategies are applied. Pre-specification of sensitivity analyses avoids data-driven selection of corresponding strategies. However, such a pre-specification may not always be possible or useful.

A widely ignored problem is the – often inadequate – reporting of loss to follow-up information in trials with time-to-event outcomes. It is essential to evaluate the censoring pattern across and between the treatment groups. If informative censoring occurs – i.e. if censoring is related to the outcomes of interest, the estimates of event rates and effect estimates are usually biased. Reviewers are encouraged to assess the consistency of loss to follow-up information, because in a recent survey of published articles reporting time-to-event outcomes it was shown that less than half of the articles reported consistent loss to follow-up information. Definitely inconsistent loss to follow-up information was presented in 15% of the articles; in about half of these a substantial change in results occurred when censored observations, which were not reported as censored in the article, were imputed (Vervölgyi et al. 2011).

Reporting bias may arise if - depending on the type of results - the results of a whole study are published (or not published) or if certain outcomes within a published study are selectively reported (or not reported): the first is typically designated as 'publication bias', while the latter is referred to as 'outcome reporting bias' (Cochrane Handbook, Dwan et al. 2008). Non-reporting of studies and outcomes is typically associated with negative results, i.e. there is a tendency not to report them at all, or to report them at a later point in time; the opposite applies to positive results (Song et al. 2010). Publication bias affects the validity of a given HTA and should therefore be considered when assessing the strength of evidence from an HTA. Outcome reporting bias might affect the internal validity of results from a given study and should therefore be evaluated as part of the risk of bias assessment. A meta-epidemiological study confirmed that outcome reporting bias is a real and serious problem, and that it has obviously been under-recognized in the past (Kirkham et al. 2010).

Besides non-reporting of outcomes, another danger is that outcome definitions (e.g. according to the operationalization of the outcome itself, the time point of assessment, the definition of cut-off points, etc.) are changed after the opening of the randomization code. Such changes obviously bear a high risk of being data-driven (Mathieu et al. 2009). Possible strategies to detect reporting bias are to search for completely unpublished studies in trial registries and to compare the original study protocol, the statistical analysis plan or entries in a trial registry with the actually reported outcomes, analyses, and data.

Other sources of bias may arise due to other reasons and in special circumstances. Examples are given below.

2.2.2. Assessment

With regard to the above mentioned types of bias and the strategies used in clinical trials to protect from such bias, the Cochrane Handbook specifies the following 7 relevant domains for the assessment of the risk of bias:

- random sequence generation (selection bias),
- allocation concealment (selection bias),
- blinding of participants and personnel (performance bias),
- blinding of outcome assessment (detection bias),
- incomplete outcome data (attrition bias),
- selective reporting (reporting bias), and

- other sources of bias, e.g. the post-hoc (potentially data-driven) definition of outcomes (e.g. the definition of the components of a composite outcome), the use of non-validated measurement instruments, or an incorrect statistical analysis.[2]

The risk of bias should be assessed on 2 levels, i.e. firstly, on a general study level, and secondly, on an outcome level. For example, possible selection and performance bias threaten the validity of the whole study, while the other types of possible bias may be outcome specific. The risk of bias may then be categorized into 3 groups:

- low risk of bias,
- high risk of bias, and
- unclear risk of bias.

A low risk of bias concerning allocation concealment, for example, can be assumed if a central allocation procedure (e.g. telephone randomization) was used in an open-label trial. Investigators are by definition kept blinded to the allocated treatment before enrolling patients into a trial when patients are to be enrolled centrally. If, however, only insufficient information with regard to the allocation procedure is provided, the risk of bias may be judged as 'unclear' or even 'high'.

If only insufficient information on specific domains is provided in the publications, in general the risk of bias remains 'unclear'. However, for the domains addressing selection bias (random sequence generation and allocation concealment), insufficient information may ultimately lead to a high risk of bias (see next paragraph), so that no 'unclear' category remains. In addition, there may be indications or no indications of 'other sources of bias', but no 'unclear' indications.

Besides the risk of bias assessment of individual domains, it may be appropriate to come to an overall conclusion across domains. There is no simple rule as to how to combine the assessments of the single domains into one overall conclusion. However, some general rules may be considered: In IQWiG reports, for example, unclear[3] allocation concealment leads to an overall high risk of bias in an open-label study. Another example: If patients are unblinded, patient-reported outcomes generally carry a high risk of bias.

The HTA assessors and readers of the present guideline are strongly encouraged to look for further details in the **Cochrane Handbook**. Very helpful support for judgment is given in the handbook (chapter 8, in particular chapter 8.5 and table 8.5.d). In addition, Annexe 2 provides a proposal for a standardized extraction sheet with instructions for completion. Furthermore, Annexe 3 provides an example of a risk of bias assessment from an IQWiG report.

2.2.3. Dealing with risk of bias

There are different strategies to deal with the risk of bias. The most stringent way is to include only outcome-specific results with a low risk of bias in a systematic review or an HTA or in meta-analyses, if a statistical pooling of the results is appropriate. This has the advantage of being as confident as possible about the findings of the evidence synthesis. However, a disadvantage of such a strategy is that the evidence base and subsequently the precision of the effect estimates will be reduced.

---

[2] There is an ongoing debate on whether studies that were stopped early for benefit carry a relevant risk of bias – despite the use of appropriate stopping rules – or not (Basler 2010, Goodman 2010). Hence, such studies are not given here as an example of 'other sources of bias'. While the Cochrane Collaboration does not regard stopping early for benefit (by using appropriate stopping rules) as an example of risk of bias (Cochrane Handbook), it is adopted in the GRADE framework. However, to be clear: a trial stopped early without appropriate stopping rules inevitably carries a high risk of bias. In addition, stopping a trial early on the basis of a certain short-term endpoint (e.g. a surrogate) may decrease the interpretability of long-term endpoints (e.g. survival) due to unblinding and crossing-over.

[3] According to IQWiG's methods (IQWiG 2011a), trials with inadequate allocation concealment are regarded to be non-randomized.

Another option is to perform sensitivity analyses according to the risk of bias categories. If estimates from study results with a high or unclear risk of bias do not substantially differ from those with a low risk of bias, it may increase confidence in the overall evidence base and allow pooling. Such an approach acknowledges that the results of a study may in fact be unbiased, despite a methodological flaw.

However, this option creates a problem: Different statistical approaches exist to assess the heterogeneity of effect estimates between study results, e.g. the I^2-statistic, or a formal statistical test on interaction. In addition, there is no general agreement on which approach is the most appropriate one, or even on thresholds defining low and high heterogeneity. Furthermore, the results of these statistical methods depend on the number of studies and the number of participants within the single studies. For this reason, it is useful to specify the way of dealing with heterogeneity in advance in the protocol for a systematic review or HTA. For example, some HTA organizations allow statistical pooling if the Cochrane Q statistics provides a p-value above 0.2. If this is the case, it is assumed that results from studies with a low or high or unclear risk of bias are not too different.

A third option is to describe the uncertainty with regard to the different levels of risk of bias, so that subsequent decisions can be made considering this uncertainty. Again, the Cochrane Handbook gives some support for interpretation. For example, a low risk of bias is interpreted as 'plausible bias unlikely to seriously alter the results'. To have a low risk of bias across studies, most information has to originate from studies with an outcome-specific low risk of bias. However, again it is not specified how 'most information' is defined. Nevertheless, it is recommended to consult the Cochrane Handbook, in particular chapter 8.7 and table 8.7a.

Some HTA agencies differentiate between the uncertainty with regard to study results (e.g. low uncertainty: RCT with a low risk of bias; moderate uncertainty: RCT with a high risk of bias; high uncertainty: non-RCT [IQWiG 2011a]) and the requirements for conclusions on the evidence base (e.g. 'proof' > 'indication' > 'hint' of the benefit or harm of an intervention [IQWiG 2011a]). For derivation of 'proof', in general results from at least 2 independent trials are required, with mostly high certainty (or low uncertainty) of results, and with effect estimates in the same direction.

Appendix 4 provides an example of how to deal with the risk of bias.

2.2.4. Systematic reviews

Systematic reviews identify, assess and summarize the evidence from one or several study types that can provide the best answer to a specific and clearly formulated question.

For systematic reviews of the effects of medical interventions, it is generally acknowledged that RCTs provide the most reliable answers. However, for other questions such as aetiology, prognosis or the qualitative description of patients' experiences, the appropriate evidence base for a systematic review will consist of other primary study types (Glasziou et al. 2004).

Systematic reviews are non-experimental studies whose methods aim to minimize systematic errors (i.e. bias) on every level of the review process (Cochrane Handbook).

In case an REA is not or not fully based on primary studies, but rather on a single systematic review or on several systematic reviews[4], it is necessary to assess whether the underlying systematic review(s) has/have only minimal methodological flaws.

Various instruments exist to assess the quality of systematic reviews (Shea et al. 2007). However, only a minority of these instruments are formally validated, widely used, and focused on methodological quality rather than on reporting quality.

One of the rare instruments which is formally validated and provides a definition of (methodological) quality is the Overview Quality Assessment Questionnaire (OQAQ), also known

[4] e.g. due to limited resources.

as the Oxman and Guyatt index (Oxman & Guyatt 1991). A further development is the AMSTAR instrument, which is based on the Oxman and Guyatt index and another checklist, and also includes additional items judged to be of actual methodological importance by experts (Shea 2007). The AMSTAR tool has also been formally validated, but is open to further improvement by advances in empirical methodological research – as acknowledged by the authors of the instrument. Both instruments focus on the (systematic) literature search, on criteria for the inclusion of primary studies, the methods for assessing the quality (i.e. internal validity) of the primary studies, and the methods for combining results. The AMSTAR tool additionally addresses possible publication bias and the handling of potential conflicts of interest of both the authors of the primary studies and those of the systematic review.

According to the Oxman and Guyatt index, systematic reviews are regarded to be of sufficient quality if they have been awarded at least 5 of 7 possible points in the overall assessment, which is performed by 2 reviewers independently of one another. No such threshold is defined for the AMSTAR Instrument and therefore should, if appropriate, be defined beforehand.

If an REA is to be performed on the basis of systematic reviews rather than on primary studies, it is strongly recommended to assess the methodological quality of the underlying reviews, either by the Oxman and Guyatt index or by the AMSTAR instrument.

3. Discussion

The certainty of results is an important criterion for the inference of conclusions on the evidence base for an REA. This certainty has both a quantitative and qualitative component. Internal validity, and hence the present guideline, deals with the qualitative component. The qualitative uncertainty of results is determined by the study design, from which evidence levels can be inferred. Non-randomized studies, for example, inevitably carry a high risk of bias. However, this uncertainty is also determined by (outcome-related) measures for further prevention or minimization of potential bias, which must be assessed depending on the study design. Such measures include, for example, the blinded assessment of outcomes, an analysis based on all included patients (potentially supported by the application of adequate replacement methods), and, if appropriate, the use of valid measurement instruments.

The recommendations of EUnetHTA for assessment of internal validity of a variety of study designs rely heavily on the latest edition (March 2011) of the Cochrane Handbook. This Handbook is regarded as representing the 'gold standard' and its use has been advocated by a number of HTA agencies active in EUnetHTA. The emphasis is on a risk of bias approach based on the following 7 principles (Higgins et al 2011):

- (1) Do not use quality scales
- (2) Focus on internal validity
- (3) Assess the risk of bias in trial results, not the quality of reporting or methodological problems that are not directly related to risk of bias
- (4) Assessments of risk of bias require judgement
- (5) Choose domains to be assessed based on a combination of theoretical and empirical considerations
- (6) Focus on risk of bias in the data as presented in the review rather than as originally reported
- (7) Report outcome-specific evaluations of risk of bias

A short rationale for each of these principles can be found in the original publication of Higgins et al. 2011, so it is not repeated here.

However, some of these principles warrant discussion. Firstly, the use of scales and checklists to assess the internal validity of studies is actively discouraged (principle 1). Nevertheless, some of the better instruments in these categories have in common that they are based on formal scale development methods. But even so, it has been increasingly acknowledged that their choice and combination is by definition arbitrary. For example, with regard to the influential Jadad scale, the developers (Jadad and Enkin 2007) themselves noticed that their scale was not the only way to assess trial quality, nor always the most appropriate one. At the same time the authors noticed that it was the most widely used scale and appeared to produce robust and valid results in an increasing number of studies. There are reasons to believe that this scale is still widely used in HTA agencies; this should change.

Occasionally, the original scale was modified to better suit local use. For example, the New Zealand Pharmaceutical Management Agency PHARMAC uses a version of the Jadad scale that is modified on the basis of the sources of bias listed by the Cochrane Handbook (PHARMAC 2005). Jadad and Enkin also indicate that the Jadad scale should not be used in isolation. It should, according to the authors, always be complemented with separate assessments of any components for which there is empirical evidence of a direct relationship with bias.

A related methodological discussion that used to play a role in the choice of instruments is the principal choice between scales and checklists, with checklists often deemed superior in the quality assessment of RCTs (e.g. Jüni et al. 2001).

Principle 3 (do not assess the quality of reporting) is also somewhat problematic, as certain guidelines (statements) for reporting have often been claimed to be a helpful tool in assessing the internal validity of studies. The most relevant examples include guidelines for reporting of RCTs (the Revised CONSORT Statement, with a couple of extensions) (e.g. Schulz et al 2010), systematic reviews and meta-analyses (the PRISMA Statement) (Moher et al. 2009), as well as guidelines for reporting observational studies (the STROBE Statement) (von Elm et al. 2007). The probability of retrieving relevant information in terms of risk of bias is higher for designs for which reporting guidelines have been in place longer. High-quality reporting, however, should not be confused with low risk of bias.

Although the Cochrane risk of bias approach can be regarded as state of the art, it should be noted that the tool is far from being perfect. In recent evaluations the inter-rater agreement on individual domains of the risk of bias tool varied between ‘slight’ and ‘substantial’ across domains (Hartling 2009, Hartling 2011). As expected, aspects requiring more judgment (e.g. selective outcome reporting) resulted in a low(er) inter-rater agreement. Nevertheless, the overall risk of bias assessment was able to differentiate treatment effect estimates (Hartling 2009). Appropriate training and clear and consistent decision rules are necessary to achieve acceptable reproducibility.

4. Conclusion

It is recommended that HTA assessors use the Cochrane risk of bias tool as an instrument to evaluate the internal validity of a study. Risk of bias has several domains and should be judged both on a study level and an outcome level. If an REA is not or not fully based on primary studies, but rather on systematic reviews (e.g. due to limited resources), it is also necessary to assess whether the underlying systematic review(s) has/have only minimal methodological flaws. It is recommended to use the Oxman and Guyatt index or the AMSTAR instrument for this purpose. Within an REA, HTA assessors should specify in advance how to deal with studies with a high or unclear risk of bias or systematic reviews with methodological shortcomings. Appropriate training and clear and consistent decision rules are necessary to achieve acceptable reproducibility when using these instruments.

Annexe 1. Methods of documentation and selection criteria

Because the Cochrane risk of bias tool can be regarded as a generally accepted standard, an extensive literature search was not conducted.

While it was not the original scope of this guideline to give recommendations on the quality assessment of systematic reviews, it nevertheless appeared to be appropriate to do so during the writing process. As far as the authors of this guideline know, only 2 instruments exist which are formally validated and are focused on methodological quality rather than on reporting quality: the Oxman and Guyatt index, and the AMSTAR instrument. Therefore an extensive literature search was not conducted.

Annexe 2. Proposal for a standardized risk of bias assessment

Criteria to assess the risk of bias in results

The extent of risk of bias in results should be estimated on the basis of the assessment of the following criteria (A: across outcomes; B: outcome-specific).

A: Aspects of the risk of bias in results at study level

A.1 Was the generation of the randomization sequence adequate?

There is no answer option 'no because in this case the trial would be classified as 'non-randomized'.

☐ **yes**: Group allocation was purely random and generation of the allocation sequence is described and is suitable (e.g. computer-generated list).

☐ **unclear**: Although the trial is described as randomized, information on the generation of the allocation sequence is missing or is not accurate enough.

if unclear, please give reasons for the classification (mandatory):

_____

_____

A.2 Allocation concealment

Procedure that ensures that the allocation of patients to the various study groups is not known to the persons who authorize the allocation or decide upon the inclusion of patients. There is no answer option 'no' because in this case the study would be classified as 'non-randomized'.

☐ **yes**: One of the following characteristics applies:

- Allocation by central, independent entity (e.g. by telephone or computer)
- Use of drugs (or drug containers) of identical appearance, numbered or coded for patients and the medical staff
- Use of a serial numbered, sealed and opaque envelope containing the group allocation.

☐ **unclear**: Information on the methods for concealing the group allocation is missing or not accurate enough.

if unclear, please give reasons for the classification (mandatory):

_____

_____

A.3 Blinding of patients and medical personnel

Patient

☐ **yes:** Patients were blinded.

☐ **unclear:** There is no information on this point.

☐ **no:** It is clear from the information that patients were not blinded.

Please give reasons for the classification (mandatory): (e.g. use of double-dummy technique)

_____

Medical personnel and other staff

☐ **yes:** Medical personnel treating the patient were blinded as to the treatment. If it is obviously impossible, e.g. in surgical procedures, to blind the primary person responsible for treatment (surgeon), it is assessed whether a suitable blinding of other staff involved in the treatment (e.g. nursing staff) took place.

☐ **unclear:** There is no information on this point.

☐ **no:** It is clear from the information that patients were not blinded.

please give reasons for the classification (mandatory):

_____

_____

A.4 Was the reporting of all relevant outcomes independent of the results?

Considerable bias can occur if the reporting of the result on an outcome depends on the nature of the result. Depending on the result, (a) reporting may be omitted, (b) the degree of detail may vary, or (c) the way of reporting may deviate from that originally planned.

Examples of a and b:

- *The primary outcome named in the sample size calculation is not/is inadequately reported in the results section.*
- *(Significant) results of not previously defined outcomes are reported.*
- *Only statistically significant results are shown with estimates and CIs.*
- *Only individual items of a score named in the methods section are reported.*

Examples of c: Selective reporting of components of the analysis:

- *Subgroups,*
- *Times/periods,*
- *Definition of outcome criteria (e.g. end-of-study value reported instead of change from baseline value; use of categorical instead of continuous values),*
- *Distance measures (e.g. odds ratio instead of risk difference),*
- *Cut-off points for dichotomization,*
- *Statistical methods.*

To estimate potential selective reporting, the following points should be considered where possible:

- *Comparison of the information in the main publication with that of other sources (trial protocol/registry report, additional publications, abstracts).*
- *Comparison of the information in the methods section with that in the results section. In particular, unless a plausible and results-independent reason is given, an actual sample size that differs markedly from that calculated is indicative of a selective termination of the study.*
  Permissible reasons are:
  - *Recognizably not results-driven, e.g. patient recruitment too slow*
  - *Sample size adjustment due to a blinded interim analysis on the basis of the scattering of the sample.*
  - *Planned interim analyses that led to a premature termination of the study.*
- *Check whether statistically non-significant results are reported in less detail.*
- *If applicable, check whether 'usual' outcomes are not reported.*

It should be noted that indications of selective reporting of a particular outcome may also apply to other outcomes and thus increase the risk of bias in results for these outcomes too. This may especially apply to cases where it is suspected that the results for individual outcomes have selectively not been reported. However, selective reporting of the results for an outcome that differs from the planned reporting does not inevitably lead to an increase in a risk of bias for other

outcomes; in this case, any selective reporting is to be entered under Point B.3 specifically for each outcome (see below).

In addition, it should be pointed out that the reporting of adverse events usually occurs in a selective manner (only increased rates/other particularities are reported); the risk of bias for other outcomes is not affected.

☐ **yes:** Selective reporting is unlikely.

☐ **unclear:** The available information does not enable this to be ascertained.

☐ **no:** The data provide indications that reporting is selective and affects the risk of bias for all relevant outcomes.

if unclear or no, please give reasons for the classification (mandatory):

_____

_____

A.5 Is the trial free from other aspects (across outcomes) that affect the risk of bias?

E.g.

- *Differing concomitant treatments between the groups outside the treatment strategies under evaluation*
- *Patient flow not transparent*
- *If planned interim analyses were performed, the following points should be observed:*
  - *The methods must be described in detail (e.g. alpha spending approach according to O'Brien Fleming, maximum sample size, planned number and time of the interim analyses).*
  - *The results (p-value, point and interval estimate) of the outcome that led to study termination should have been adjusted (otherwise, if applicable, to be carried out post hoc by the responsible HTA agency).*
  - *Adjustment should then also take place if the maximum sample size was reached.*
  - *If other outcomes are correlated with the outcome that led to study termination, these should also be adequately adjusted.*

☐ **yes**

☐ **no**

if no, please give reasons for the classification (mandatory):

_____

_____

Classification of the risk of bias in results at study level:

Classification of the risk of bias in results takes account of the individual assessments of the previous Points A.1 to A.5. A relevant bias means that the basic conclusions from the results would be changed if the biased aspects were corrected.

☐ **low**: There is a high probability that the possibility of bias in results caused by these aspects across outcomes can be ruled out.

☐ **high**: The results are possibly subject to relevant bias.

If high, please give reasons for the classification (mandatory):

_____

_____

B: Aspects of the risk of bias in results by outcome

The following Points B.1 to B.4 are used to estimate the outcome-specific aspects for the extent of possible bias in results. These points should generally be assessed for each relevant outcome separately (if applicable, several outcomes can be assessed together e.g. outcomes regarding adverse events).

Outcome: _____

B.1 Was the outcome assessor blinded?

Determine whether the person who assessed the outcome was blinded in relation to the treatment. In some cases, blinding may also be required for the results of other outcomes (e.g. typical adverse events), if knowledge of these results potentially indicates the type of administered treatment and thus may lead to unblinding.

☐ **yes:** The outcome was assessed in a blinded manner.

☐ **unclear:** There is no information on this point.

☐ **no:** It is clear from the information that no blinded assessment took place.

Please give reasons for the classification (mandatory)

_____

_____

B.2 Was the ITT (intention-to-treat) principle appropriately implemented?

Lost to follow-up patients are those in whom the outcome criteria could not be fully assessed right up to the end of the study (e.g. because a patient withdrew his/her consent). Protocol violators include patients who did not complete the allocated treatment according to the protocol (e.g. those who stopped or changed treatment or who took non-permitted concomitant medication). It should be noted that terms such as lost to follow-up and protocol violators are, however, sometimes defined very differently in publications. In addition, terms such as drop-outs, withdrawals etc. should be avoided as far as possible in this extraction form or precisely defined. If such patients occur in a study, they must be adequately and fully described (reasons for discontinuation, frequency and patient characteristics per group) or appropriately considered in the statistical analysis (generally ITT analysis). In an ITT analysis all randomized patients are analysed according to their group allocation (where applicable, missing values for the outcome criteria in lost to follow-up patients must be replaced in a suitable way). It should be noted that the term ITT is not always used in this strict sense in publications. Often only the randomized patients who at least began the treatment and for whom at least one post-baseline value has been recorded (full analysis set) are analysed. In justified cases, this procedure is guideline-compliant, but an assessment of potential bias should be conducted, particularly in non-blinded studies. In equivalence and non-inferiority studies, it is especially important that such patients are described very precisely and the methods for taking account of these patients are shown in a transparent manner.

☐ **yes:** One of the following characteristics applies:

- According to the publication, no protocol violators or lost to follow-up patients occurred in relevant numbers (if applicable, to be defined in the project, e.g. non-consideration rate in the analysis <5%), and there is no evidence (e.g. discrepant patient numbers in flow chart and results table) to doubt this.
- The protocol violators and lost to follow-up patients are to be described in such detail (nature, frequency and characteristics per group) that their possible influence on the results can be estimated (independent analysis possible).
- The strategy to take account of protocol violators and lost to follow-up patients (including the replacement of missing values, choice of outcome criteria, statistical methods) is logically designed (does not bias the effects in favour of the treatment being evaluated).

☐ **unclear**: Due to inadequate reporting, the proper handling of protocol violators and lost to follow-up patients cannot be assessed.

☐ **no**: None of the 3 characteristics named under 'yes' applies.

if unclear or no, please give reasons for the classification (mandatory):

_____

_____

B.3 Was the reporting of this outcome independent of the results?

Note the advice on Point A.4.

☐ **yes:** Selective reporting is unlikely.

☐ **unclear:** No evaluation possible from the information available.

☐ **no:** The data provide indications of selective reporting.

If unclear or no, please give reasons for the classification (mandatory):

_____

_____

B.4 Is the trial free from other (outcome-specific) aspects that affect the risk of bias?

E.g.

- *Relevant data inconsistencies within or between the publications (and, if applicable, other documents).*
- *Implausible information*
- *Use of inadequate statistical methods*

☐ **yes**

☐ **no**

If no, please give reasons for the classification (mandatory):

_____

_____

Classification of the risk of bias in the results for this outcome:

The risk of bias is classified in conjunction with the individual assessments of the previous outcome-specific Points B.1 to B4. and the classification of the risk of bias at study level. If the risk of bias across outcomes was rated as 'high', the risk of bias for the outcome is also to be rated as 'high'. A relevant bias means that the basic conclusions from the results would be changed if the biased aspects were corrected.

☐ **low**: There is a high probability that the results for this outcome are not subject to relevant bias caused by outcome-specific aspects and aspects across outcomes.

☐ **high**: The results are possibly subject to relevant bias.

If high, please give reasons for the classification (mandatory):

Annexe 3. Example of a risk of bias assessment

The example is taken from an IQWiG benefit assessment of ezetimibe for hypercholesterolaemia (IQWiG 2011b). In brief, ezetimibe (in mono- or combination therapy) was compared to treatment with placebo or other lipid-lowering drugs, as well as to non-drug treatment options in patients with primary hypercholesterolaemia. The focus of the assessment was on patient-relevant outcomes. The assessment was based on RCTs with a follow-up of at least 12 months. Ezetimibe had to be administered according to the approval status in Germany.

Two RCTs were included in this assessment: the 24-month ENHANCE study (e.g. Kastelein 2008) and the 14-month ARBITER-6-HALTS study (e.g. Taylor 2009), which included a total of 720 and 363 patients respectively. On the basis of therapy with statins, the trials investigated the additional administration of ezetimibe versus placebo (ENHANCE) or niacin (ARBITER-6-HALTS).

ENHANCE was a double-blind study, and randomization was based on computer-generated codes provided to the clinical centres by a central randomization service (→ adequate generation of randomization sequence and allocation concealment). While it was not explicitly stated in the publications that the placebos were identical to the study drug with regard to outward appearance and taste, it was nevertheless assumed that real double blinding was achieved (because – among other things – in the registry entry [ClinicalTrials.gov identifier NCT00552097] it is mentioned that subjects, investigators and outcome assessors were blinded). Selective outcome reporting was as far as possible excluded (the clinical study report was provided by the pharmaceutical company), and no other aspects were identified which carried a risk of bias. Hence, the risk of bias on the study level was judged as ‘low’ (table 1). It should be noted that in IQWiG reports, incomplete reporting of outcome data is not assessed on the study level, but only on the outcome level.

In contrast, ARBITER-6-HALTS was an open-label study. Because no details were provided in the publications on how allocation concealment was achieved (allocation concealment unclear), the risk of bias on the study level was judged as ‘high’ (table 1).

The risk of bias assessment on the outcome level is shown in table 2. The risk of bias on the outcome level was assessed as ‘high’ for most outcomes of the ARBITER-6-HALTS study, mainly because a high number of dropouts were not included in the analyses. In addition, there was an indication of selective exclusion of patients from the analyses. Furthermore, the results for one component of a combined endpoint were not reported, and for another outcome (health-related quality of life) no detailed information with regard to the assessment of the outcome and the results was provided. Both latter deficiencies represent examples of ‘other sources of bias’.

Table 1: Risk of bias – study level (table adapted from IQWiG 2011b)*

Trial	Adequate generation of randomization sequence	Adequate allocation concealment	Blinding		Selective outcome reporting unlikely	No other aspects which increase the risk of bias	Risk of bias – study level
			Patient	Treating Physician			
ENHANCE	yes	yes	yes	yes	yes	yes	low
ARBITER-6-HALTS	yes	unclear[a]	no	no	yes	yes	high[a]
a: No details were provided on how allocation concealment was achieved (in the open-label trial).							

*The item 'incomplete reporting of outcome data' is missing here because it is assessed only on an outcome level in IQWiG reports.

Table 2: Risk of bias – outcome level (summarized assessment, table adapted from IQWiG 2011b)

Outcome ►			Mortality			Morbidity					Adverse events		
Trial▼	Study level	Overall survival	Cardiac	Cerebral	Non-cardiac and non-cerebral	Cardiac	Cerebral	Non-cardiac and non-cerebral	Combined endpoint (cardiac mortality and morbidity)	Health-related quality of life	AE	SAE	Treatment discont. due to AEs
ENHANCE	low	low	low	low	low	low	low	–	low	–	low	low	low
ARBITER-6-HALTS	high	high[a,b]	high[a,c]	–	–	high[a,c]	–	–	high[a,c,d]	high[a,e]	–	–	high[a]

a: High risk of bias on study level.
b: It is unclear whether dropouts were included in the analysis (incomplete reporting of outcome data).
c: High number of dropouts not included in the analysis (incomplete reporting of outcome data); in addition, large difference in patients not included in the analysis between treatment groups. (9 [5%] in the ezetimibe group and 27 [14%] in the niacin group).
d: The results for one component of the combined endpoint were not reported (selective outcome reporting).
e: No detailed information with regard to the assessment of the endpoint and results; it was only noted that no statistically significant difference between the treatment groups was observed..

–: Endpoint was not assessed/reported.

discont.: discontinuation; SAE: severe adverse event; AE: adverse event

Table 3: Risk of bias – outcome level (overall and vascular mortality, table adapted from IQWiG 2011b)

Outcome Trial	**Risk of bias – study level**	**Blinding – outcome assessors**	**ITT principle adequately realized**	**Selective outcome reporting unlikely**	**No other aspects according to risk of bias**	**Risk of bias – outcome level**
Overall survival						
ENHANCE	low	yes	yes	yes	yes	low
ARBITER-6-HALTS	high	no	unclear[a]	yes	yes	high[a,b]
Vascular mortality (cardiac)						
ENHANCE	Low	yes	yes	yes	yes	low
ARBITER-6-HALTS	high	no[c]	no[d]	yes	yes	high[b,d]
Vascular mortality (cerebral)						
ENHANCE	low	yes	yes	yes	yes	low
ARBITER-6-HALTS	high	–	–	–	–	–
Vascular mortality (non-cardiac and non-cerebral)						
ENHANCE	low	yes	yes	yes	yes	low
ARBITER-6-HALTS	high	–	–	–	–	–

a: It is unclear whether dropouts were included in the analysis
b: High risk of bias on study level.
c: Open-label trial. Clinical endpoints were adjudicated by an independent data advisory committee who were unaware of the treatment assignments. However, it is assumed that the primary documentation of cardiovascular events was unblinded.
d: High number of dropouts not included in the analysis; in addition, large difference in patients not included in the analysis between treatment groups. (9 [5%] in the ezetimibe group and 27 [14 %] in the niacin group).

–: Endpoint was not assessed/reported.

Annexe 4. Example of dealing with risk of bias

The example is taken from an IQWiG benefit assessment of rivastigmine patch (10 cm²) and other drugs for Alzheimer's disease (IQWiG 2012). In brief, rivastigmine patch (10 cm²) was compared to treatment with placebo or other drugs for Alzheimer's disease, as well as to non-drug treatment options in patients with mild to moderate Alzheimer's disease. The focus of the assessment was on patient-relevant outcomes. The assessment was based on RCTs with a follow-up of at least 16 weeks. Rivastigmine patch had to be administered according to the approval status in Germany.

Two RCTs were included in this assessment for the comparison against placebo: the 24-week, multi-national IDEAL study (e.g. Winblad 2007) and the Japanese study D1301 (unpublished, Novartis 2011[5]), which included a total of 892 and 859 patients respectively[6]. The following discussion is based on the outcome 'cognitive function'.

Cognitive function was measured in both studies by the Alzheimer's Disease Assessment Scale Cognitive subscale (ADAS-cog) and amongst others operationalized as 'non-reponse', i.e. an improvement of less than 4 points on the ADAS-cog scale. The cut-off point of 4 is an established threshold for non-response on the ADAS-cog scale. The risk of bias with regard to this outcome was judged to be high for the IDEAL study and low for the study D1301. The reason for the high risk of bias judgment for the IDEAL study was a major difference (> 5%) in the proportion of non-analysed patients between the rivastigmine patch and placebo group.

A meta-analysis of both trials showed a statistically significant effect in favour of rivastigmine (figure 1, $p = 0.023$) with no statistical heterogeneity ($I^2 = 0\%$). According to the IQWiG methods, a statistically significant effect of a meta-analysis of at least 2 studies with an outcome-specific low risk of bias is in general necessary to derive 'proof' of a benefit (highest level of certainty). Therefore, in the first instance only an 'indication' (middle level of certainty) of a benefit was acknowledged, because one of the 2 studies had a high risk of bias with regard to cognitive function.

In addition 2 sensitivity analyses were performed to evaluate the robustness of the results. In the first sensitivity analysis all patients who were not included in the primary analysis were regarded as non-responders (figure 2). This analysis also showed a statistically significant effect in favour of rivastigmine ($p = 0.046$). In the second sensitivity analysis all patients who were not observed until the last planned follow-up were regarded as non-responders[7] (figure 3). Here, the statistically significant effect disappeared ($p = 0.124$). Therefore, it was concluded that the effect was not sufficiently robust to derive 'proof' of a benefit, so that the results were still regarded as an 'indication' of a benefit[8].

[5] A study report was provided by Novartis for the assessment.
[6] However, for the comparison of rivastigmine patch (10 cm²) with placebo, 595 and 575 patients were included respectively.
[7] The primary analysis used a last-observation-carried-forward (LOCF) strategy.
[8] Because there was also an indication for an effect modification by older age (< / ≥ 75 years) the conclusions were drawn separately for patients < 75 years of age ('indication' of a benefit) and ≥ 75 years ('hint' of a benefit).

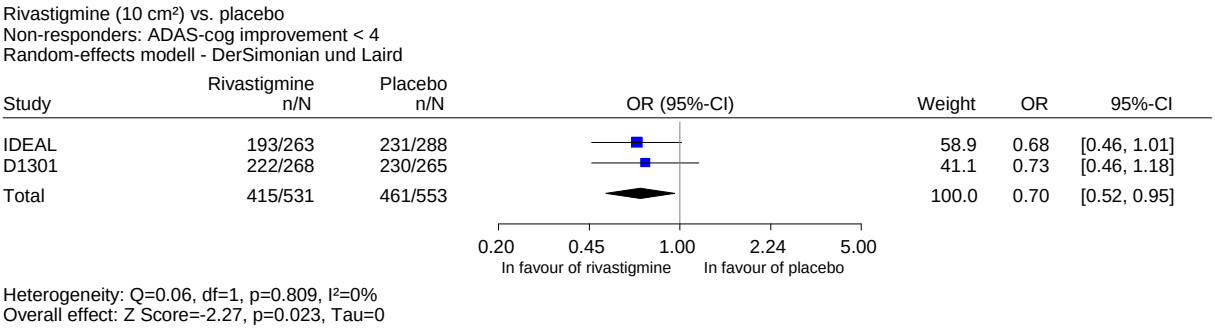

Figure 1: Meta-analysis of rivastigmine patch (10 cm²) vs. placebo, ADAS-cog non-response. OR = odds ratio; CI = confidence interval.

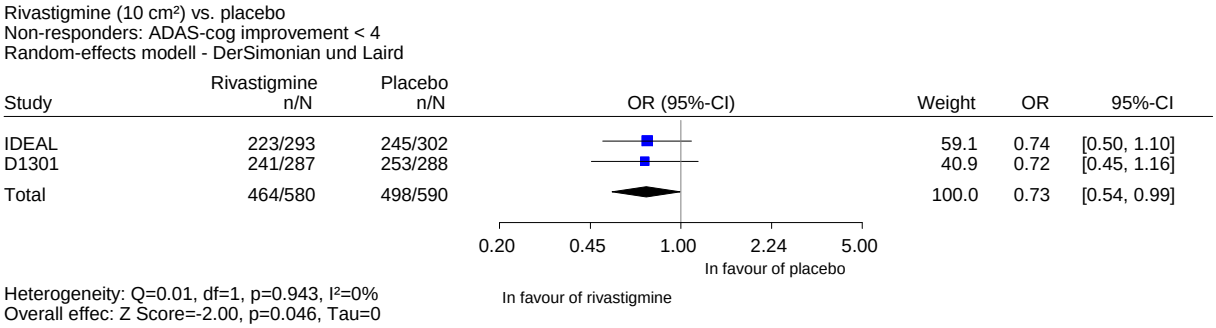

Figure 2: Meta-analysis of rivastigmine patch (10 cm²) vs. placebo, ADAS-cog non-response. Sensitivity analysis 1: Patients who were not included in the primary analysis were regarded as non-responders. OR = odds ratio; CI = confidence interval.

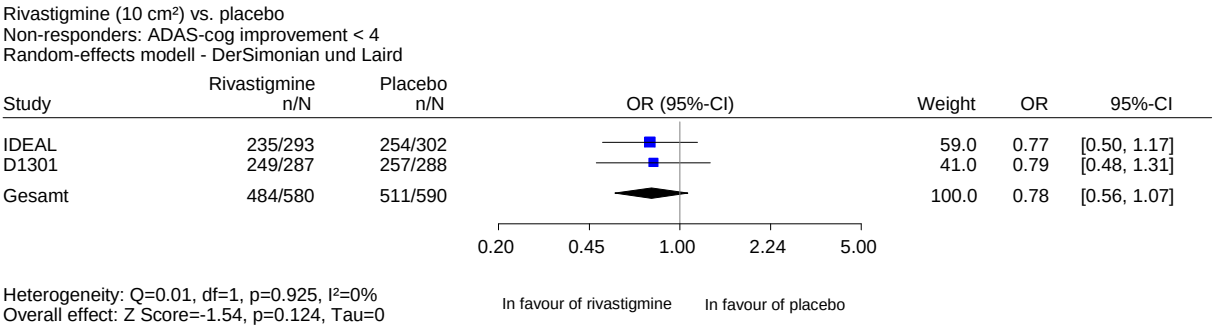

Figure 3: Meta-analysis of rivastigmine patch (10 cm²) vs. placebo, ADAS-cog non-response. Sensitivity analysis 2: Patients who were lost to follow-up were regarded as non-responders. OR = odds ratio; CI = confidence interval.

Annexe 5. Bibliography

1) Bassler D, Briel M, Montori VM, Lane M, Glasziou P, et al. Stopping randomized trials early for benefit and estimation of treatment effects: systematic review and meta-regression analysis. JAMA 2010; 303: 1180-1187.

2) Dwan K, Altman DG, Arnaiz JA, Bloom J, Chan AW, Cronin E, Decullier E, Easterbrook PJ, Von Elm E, Gamble C, Ghersi D, Ioannidis JP, Simes J, Williamson PR: Systematic review of the empirical evidence of study publication bias and outcome reporting bias. PLoS ONE 2008, 3:e3081.

3) [EMA] European Medicines Agency. Reflection paper on the regulatory guidance for the use of health related quality of life (HRQL) measures in the evaluation of medicinal products [online]. 27.07.2005. Available at URL: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003637.pdf (accessed: 27.12.2012).

4) [EMA] European Medicines Agency. Appendix 1 to the guideline on the evaluation of anticancer medicinal products in man. Methodological consideration for using progression-free survival (PFS) or disease-free survival (DFS) in confirmatory trials [online]. 15.12.2011. Available at URL: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2011/12/WC500119965.pdf (accessed: 27.12.2012).

5) [EUnetHTA]. European network on HTA. Common questions. What is Health Technology Assessment (HTA) [online]. Available at URL: http://www.eunethta.eu/about-us/faq#t287n73 (accessed: 27.12.2012).

6) [FDA] Food and Drug Administration. Guidance for industry. Patient-reported outcome measures: Use in medical product development to support labeling claims [online]. 12.2009. Available at URL: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM193282.pdf (accessed: 27.12.2012).

7) Glasziou PP, Vandenbroucke JP, Chalmers I. Assessing the quality of research. BMJ 2004; 328: 39-41.

8) Goodman S, Berry D, Wittes J. Bias and trials stopped early for benefit. JAMA 2010; 304: 157.

9) Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, Montori V, Akl EA, Djulbegovic B, Falck-Ytter Y, Norris SL, Williams JW Jr, Atkins D, Meerpohl J, Schünemann HJ. GRADE guidelines: 4. Rating the quality of evidence--study limitations (risk of bias). J Clin Epidemiol. 2011; 64: 407-415.

10) Hartling L, Ospina M, Liang Y, Dryden DM, Hooton N, Krebs Seida J, Klassen TP. Risk of bias versus quality assessment of randomised controlled trials: cross sectional study.. BMJ 2009; 339 :b4012.

11) Hartling L, Bond K, Vandermeer B, Seida J, Dryden DM, Rowe BH. Applying the risk of bias tool in a systematic review of combination long-acting beta-agonists and inhaled corticosteroids for persistent asthma. PLoS One. 2011; 6(2) :e17242.

12) Herbison P, Hay-Smith J, Gillespie WJ. Different methods of allocation to groups in randomized trials are associated with different levels of bias. A meta-epidemiological study. J Clin Epidemiol 2011; 64: 1070-5.

13) Higgins JP, Green S (editors). Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011] [online]. The Cochrane Collaboration, 2011. Available at URL: www.cochrane-handbook.org (accessed: 27.12.2012).

14) Higgins JP, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, Savovic J, Schulz KF, Weeks L, Sterne JA; Cochrane Bias Methods Group; Cochrane Statistical Methods Group. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ 2011; 343: d5928.

15) Hróbjartsson A, Thomsen AS, Emanuelsson F, Tendal B, Hilden J, Boutron I, Ravaud P, Brorson S. Observer bias in randomised clinical trials with binary outcomes: systematic review of trials with both blinded and non-blinded outcome assessors. BMJ 2012; 344: e1119.

16) IQWiG 2011a. General Methods - Version 4.0 [online]. 23.09.2011. Available at URL: https://www.iqwig.de/download/General_Methods_4-0.pdf (accessed: 27.12.2012).

17) IQWiG 2011b. Ezetimib bei Hypercholesterinämie. IQWiG-Berichte – Jahr: 2011 Nr. 90 [online]. 18.07.2011. Available at URL: https://www.iqwig.de/download/A10-02_Abschlussbericht_Ezetimib_bei_Hypercholesterinaemie.pdf (accessed: 27.12.2012).

18) IQWiG 2012. Cholinesterasehemmer bei Alzheimer Demenz: Ergänzungsauftrag Rivastigmin-Pflaster und Galantamin. IQWiG-Berichte – Jahr: 2012 Nr. 118 [online]. 03.02.2012. Available at URL: https://www.iqwig.de/download/A09-05_Abschlussbericht_Cholinesterasehemmer_Ergaenzungsauftrag_Rivastigmin_Pflaster_Galantamin.pdf (accessed: 27.12.2012).

19) Jadad AR, Murray WE. Randomized controlled trials: questions, answers and musings. Malden: Blackwell; 2007.

20) Jüni P, Altman DG, Egger M. Assessing the quality of controlled clinical trials. BMJ 2001; 323: 42-46.

21) Kastelein JJ, Akdim F, Stroes ES, Zwinderman AH, Bots ML, Stalenhoef AF, Visseren FL, Sijbrands EJ, Trip MD, Stein EA, Gaudet D, Duivenvoorden R, Veltri EP, Marais AD, de Groot E; ENHANCE Investigators. Simvastatin with or without ezetimibe in familial hypercholesterolemia. N Engl J Med 2008; 358:1431-1443.

22) Kirkham JJ, Dwan KM, Altman DG, Gamble C, Dodd S, Smyth R, Williamson PR. The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. BMJ 2010; 340: c365.

23) Lange S. The all randomized/full analysis set (ICH E9): may patients be excluded from the analysis? Drug Inf J 2001; 35(3): 881-891.

24) Mathieu S, Boutron I, Moher D, Altman DG, Ravaud P. Comparison of registered and published primary outcomes in randomized controlled trials. JAMA 2009; 302: 977-84.

25) Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. BMJ 2009; 339: b2535.

26) Novartis Pharmaceuticals. Efficacy and safety of rivastigmine transdermal patch in patients with mild to moderate Alzheimer's disease [online]. In: ClinicalTrials.gov. 27.09.2011. Available at URL: http://ClinicalTrials.gov/show/NCT00423085 (accessed: 27.12.2012).

27) Oxman AD, Guyatt GH. Validation of an index of the quality of review articles. J Clin Epidemiol 1991; 44: 1271-1278.

28) PHARMAC 2005. Recommended methods to derive clinical inputs for proposals to PHARMAC. Version 1B, New Zealand Pharmaceutical Management Agency Ltd (PHARMAC), 2005.

29) Pharmaceutical Forum. Core principles on relative effectiveness. Available at URL: http://ec.europa.eu/enterprise/sectors/healthcare/files/docs/rea_principles_en.pdf (accessed: 27.12.2012).

30) Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA 1995; 273: 408-12.

31) Schulz KF, Altman DG, Moher D, for the CONSORT Group. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials.BMJ 2010; 340: c332.

32) Shea BJ, Grimshaw JM, Wells GA, Boers M, Andersson N, Hamel C, Porter AC, Tugwell P, Moher D, Bouter LM. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. BMC Med Res Methodol. 2007 Feb 15; 7: 10.

33) Song F, Parekh S, Hooper L, Loke YK, Ryder J, Sutton AJ, Hing C, Kwok CS, Pang C, Harvey I. Dissemination and publication of research findings: an updated review of related biases. Health Technol Assess 2010; 14(8): iii, ix-xi, 1-193.

34) Taylor AJ, Villines TC, Stanek EJ, Devine PJ, Griffen L, Miller M, Weissman NJ, Turco M. Extended-release niacin or ezetimibe and carotid intima-media thickness. N Engl J Med 2009; 361: 2113-2122.

35) Unnebrink K, Windeler J. Intention-to-treat: methods for dealing with missing values in clinical trials of progressively deteriorating diseases. Statist Med 2001; 20: 3931-3946.

36) Vervölgyi E, Kromp M, Skipka G, Bender R, Kaiser T. Reporting of loss to follow-up information in randomised controlled trials with time-to-event outcomes: a literature survey. BMC Med Res Methodol 2011 Sep 21; 11: 130.

37) von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. PLOS Medicine 2007; 4: e296.

38) Winblad B, Grossberg G, Frolich L, Farlow M, Zechner S, Nagel J et al. IDEAL: a 6-month, double-blind, placebo-controlled study of the first skin patch for Alzheimer disease. Neurology 2007; 69(4 Suppl 1): S14-S22.

39) Wood L, Egger M, Gluud LL, Schulz KF, Jüni P, Altman DG et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. BMJ 2008; 336: 601-605.