

KCE PROCESS NOTE: RESEARCH METHODS FOR GUIDELINES – UPDATE 2025

UPDATE 2025



Title:	KCE process note: Research Methods for guidelines – update 2025
Authors:	Vicky Jespers (KCE), Leen Verleye (KCE), Elena Costa (KCE), Joan Vlayen (Sint-Trudo Ziekenhuis), Thomas Struyf (Struyf Diagnostic Expertise)
Project manager:	Els Van Bruystegem (KCE)
External reviewers:	Trudy Bekkering (Katholieke universiteit Leuven (KU Leuven)), Simon Van Cauwenbergh (Werkgroep Ontwikkeling Richtlijnen Eerste Lijn (WOREL))
Internal reviewers	Roos Leroy (KCE), Ann Van den Bruel (KCE)
Acknowledgements:	The following persons contributed to previous (online) versions: Nadia Benahmed (KCE), Patrice Chalon (KCE), Anja Desomer (KCE), Lotty Hooft (Cochrane Netherlands), (Universitair Medisch Centrum, Utrecht, the Netherlands (UMC Utrecht), (Universiteit Utrecht, the Netherlands)), Pascal Jonckheer (KCE), Miranda Langendam (Amsterdam Public Health Research Institute, the Netherlands), Jo Robays ((Federaal agentschap voor geneesmiddelen en gezondheidsproducten – Agence fédérale des médicaments et des produits de santé (FAGG – AFMPS)), Rob Scholten (Julius Center for Health Sciences and Primary Care, the Netherlands), (UMC Utrecht, the Netherlands) (Universiteit Utrecht, the Netherlands)), Nancy Thiry (KCE), Mariska Tuut (Care and Public Health Research Institute, the Netherlands (CAPHRI)), Maastricht University, the Netherlands), Leen Verleye (KCE), Joan Vlayen (Sint-Trudo Ziekenhuis)
Reported interests KCE members:	KCE is a federal institution funded by INAMI/RIZIV, by the Federal Public Service of Health, Food chain Safety and Environment, and by the Federal Public Service of Social Security. KCE's mission is to advise policymakers on decisions relating to health care and health insurance on the basis of scientific and objective research. It is expected to identify and shed light on the best possible solutions, in the context of an accessible, high-quality health care system with due regard for growing demand and budgetary constraints. KCE has no interest in companies (commercial or non-commercial, i.e. hospitals and universities), associations (e.g. professional associations, unions), individuals or organisations (e.g. lobby groups) that could be positively or negatively affected (financially or in any other way) by the implementation of the recommendations. All experts involved in the writing of the report or the peer-review process completed a declaration of interest form. Information on potential conflicts of interest is published in the colophon of this report. All members of the KCE Expert Team make yearly declarations of interest and further details of these are available upon request.
Layout:	Ine Verhulst

Disclaimer:

- **The external reviewers critically evaluated a (preliminary) version of the scientific report. They did not co-author the scientific report and did not necessarily agree with its content.**
- **Finally, this report has been approved by common assent by the Executive Board.**
- **Only the KCE is responsible for errors or omissions that could persist. The policy recommendations are also under the full responsibility of the KCE.**

Publication date:	06 January 2026
Project number:	2024-70-METHOD
Domain:	METHOD
MeSH:	Guidelines, methodology, research methods
NLM Classification:	84.3



Language: English

Format: Adobe® PDF™ (A4)

Legal depot: D/2025/10.273/51

ISSN: 2466-6459

Copyright: KCE reports are published under a “by/nc/nd” Creative Commons Licence
<http://kce.fgov.be/content/about-copyrights-for-kce-publications>.



How to refer to this document?

Jespers V, Verleye L, Costa E, Vlayen J, Struyf T. KCE process note: Research Methods for guidelines – update 2025. Method. Brussels: Belgian Health Care Knowledge Centre (KCE). 2025. KCE Reports 411.
<https://doi.org/10.57598/R411C>.

This document is available on the website of the Belgian Health Care Knowledge Centre.



■ TABLE OF CONTENTS

1	INTRODUCTION	9
1.1	RATIONALE.....	9
1.2	AIM AND METHODS.....	10
2	GENERAL ASPECTS OF GUIDELINE DEVELOPMENT	10
2.1	THE GUIDELINE DEVELOPMENT GROUP.....	10
2.2	STAKEHOLDERS.....	10
2.3	PATIENT INVOLVEMENT	11
2.4	CONFLICT OF INTEREST AND KCE POLICY (UPDATE 2025)	11
2.5	DECISION-MAKING PROCESS (UPDATE 2025)	12
3	DEFINING THE SCOPE AND DEVELOPING REVIEW QUESTIONS.....	13
3.1	SEARCH FOR EXISTING GUIDELINES	13
3.2	PRIORITISATION OF RESEARCH QUESTIONS	13
3.3	PICOS AND OUTCOMES	13
3.4	THE GRADE ADAPT AND ADOPT APPROACH (UPDATE 2025)	14
3.4.1	Adapting or adopting recommendations from a source guideline.....	14
3.4.2	Identifying a relevant source guideline of quality	14
3.4.3	Deciding on adapt and adopt and next steps	14
3.5	SEARCH FOR EVIDENCE.....	17
3.5.1	Selection of systematic reviews (Update 2025)	18
3.5.2	Selection of primary studies	19
4	CRITICAL APPRAISAL OF THE LITERATURE	20
4.1	INTRODUCTION (UPDATE 2025)	20
4.2	SYSTEMATIC REVIEWS	20
4.2.1	AMSTAR-2	20
4.2.2	ROBIS (Update 2025)	21
4.2.3	ROB-ME (Update 2025)	21
4.3	RANDOMISED CONTROLLED TRIALS FOR INTERVENTIONS (UPDATE 2025).....	22
4.3.1	RoB 1 tool.....	22
4.3.2	RoB 2 tool.....	23
4.3.3	ROBUST-RCT	24
4.3.4	Other tools	24
4.4	DIAGNOSTIC ACCURACY STUDIES (UPDATE 2025)	24
4.4.1	Introduction to QUADAS-2	24
4.4.2	Domain 1: Participant Selection	26



4.4.3	Domain 2: Index test	31
4.4.4	Domain 3: Reference Standard.....	34
4.4.5	Domain 4: Flow and Timing.....	37
4.4.6	Visualising and presenting QUADAS-2 results	40
4.4.7	QUADAS-C.....	42
4.4.8	QUAPAS.....	43
4.5	OBSERVATIONAL STUDIES (UPDATE 2025).....	44
4.5.1	ROBINS-I.....	45
4.5.2	ROBINS-E	45
4.5.3	Newcastle-Ottawa Scale	46
4.6	RISK OF BIAS IN NETWORK META-ANALYSIS	46
5	GRADE FOR INTERVENTIONS	46
5.1	INTRODUCTION	46
5.1.1	What is GRADE? (Update 2025).....	46
5.1.2	Applicability of GRADE (Update2025)	47
5.1.3	Steps in the GRADE process	48
5.2	FRAMING THE QUESTION	50
5.3	DECIDING ON THE IMPORTANCE OF OUTCOMES.....	50
5.3.1	General approach.....	50
5.3.2	Perspective of outcomes	51
5.3.3	Before and after literature review	51
5.3.4	Implications of the classification	52
5.3.5	Use of surrogates	52
5.3.6	Decision threshold and minimal important difference (Update 2025)	52
5.3.7	Outcome 'as yet undiscovered toxicity' for new drugs	53
5.4	GRADE EVIDENCE PROFILES AND SUMMARY OF FINDINGS TABLES	53
5.5	RATING THE CERTAINTY OF EVIDENCE (UPDATE 2025).....	54
5.5.1	Introduction (Update 2025).....	54
5.5.2	Study limitations, Risk of bias (Update 2025)	56
5.5.3	Inconsistency (Update 2025).....	58
5.5.4	Indirectness (Update 2025).....	63
5.5.5	Imprecision (Update 2025)	65
5.5.6	Publication bias (Update 2025)	70
5.6	GRADE FOR NETWORK META-ANALYSIS (UPDATE 2025).....	72
5.6.1	Introduction	72
5.6.2	Rating the certainty of the direct (step 1) and indirect (step 2) estimates.....	73



5.6.3	Rating the certainty of the network estimates (step 3).....	73
5.6.4	Glossary network meta-analysis and further reading.....	74
5.7	SUMMARY WITHOUT META-ANALYSIS (SWIM)	75
5.7.1	Summarising effect estimates	75
5.7.2	Combining P values	75
5.7.3	Vote counting based on the direction of effect.....	76
5.7.4	Reporting of a synthesis without meta-analysis	76
6	GRADE FOR DIAGNOSTIC TESTS AND TEST-TREATMENT MANAGEMENT STRATEGIES.....	77
6.1	INTRODUCTION (UPDATE 2025)	77
6.2	FRAMING THE QUESTION	78
6.2.1	The purpose of a test or diagnostic strategy	78
6.2.2	The format of the question	78
6.3	GRADE DTA EVIDENCE PROFILES AND SUMMARY OF FINDINGS TABLES (UPDATE 2025).....	80
6.4	DECIDING ON THE IMPORTANCE OF OUTCOMES.....	82
6.4.1	Defining outcomes in a GRADE diagnostic framework.....	82
6.4.2	Outcome-specific decision thresholds (Update 2025).....	83
6.5	PRESENTING AND VISUALISING DTA RESULTS (UPDATE 2025).....	83
6.5.1	Text and Summary Tables	84
6.5.2	Forest Plots	84
6.5.3	Receiver Operating Characteristic (ROC) plots	86
6.5.4	Dumbbell plots.....	87
6.5.5	Probability plot.....	90
6.5.6	When to use which plot?	92
6.6	META-ANALYSIS OF DTA STUDIES (UPDATE 2025).....	93
6.6.1	Introduction.....	93
6.6.2	Bivariate random-effects meta-analysis	93
6.6.3	Hierarchical Summary ROC (HSROC) Curves.....	97
6.6.4	Analysis tools for meta-analysis in DTA reviews.....	99
6.7	RATING THE CERTAINTY OF EVIDENCE (UPDATE 2025).....	101
6.7.1	Study limitations, risk of bias	102
6.7.2	Indirectness	103
6.7.3	Inconsistency, heterogeneity.....	103
6.7.4	Imprecision	106
6.7.1	Publication bias in DTA reviews	107
7	RECOMMENDATIONS	109



7.1	EVIDENCE TO DECISION FRAMEWORK (UPDATE 2025)	109
7.2	WORDING THE STRENGTH OF A RECOMMENDATION (UPDATE 2025)	111
7.3	FORMULATING GOOD PRACTICE STATEMENTS (UPDATE 2025)	112
8	REVIEW AND VALIDATION	113
8.1	STAKEHOLDER REVIEW	113
8.2	VALIDATION BY EXTERNAL EXPERTS	113
9	DISSEMINATION	113
10	MANUAL KCE - VALIDATED ASSESSMENT TOOLS	113
10.1	AGREE II – GUIDELINES	114
10.2	AMSTAR 2 – SYSTEMATIC REVIEWS	115
10.3	ROBIS – SYSTEMATIC REVIEWS	119
10.4	QUADAS-2 – DIAGNOSTIC STUDIES	122
10.5	ROB-1 – RCTS	123
10.6	ROBINS-I – NON-RANDOMISED STUDIES OF INTERVENTIONS	125

LIST OF FIGURES

Figure 1 – Standard KCE GCP report process	9
Figure 2 – The GRADE ADAPT and ADOPT approach	15
Figure 3 – Overview of the Core GRADE approach	49
Figure 4 – The certainty of evidence by the GRADE approach	55
Figure 5 – Constant relative risk for varying baseline risk resulting in differing absolute risks	59
Figure 6 – Judging inconsistency with Core GRADE	60
Figure 7 – Optimal information size by control group event rate	68
Figure 8 – Total number of events needed by control group event rate	69
Figure 9 – Flow chart depicting process of deciding whether to rate down certainty of evidence for publication bias	71
Figure 10 – Flow for rating the certainty of a network estimate	73
Figure 11 – Generic study designs to evaluate the impact of testing	77
Figure 12 – Example Forest plot of C-reactive protein (CRP) for the diagnosis of pneumonia, ordered by: CRP threshold, setting, and sensitivity	84
Figure 13 – ROC plot of C-reactive protein (CRP) for the diagnosis of pneumonia: same data as the Forest Plot example	87
Figure 14 – Dumbbell plot of blood tests for the diagnosis of sepsis and pneumonia in older adults visiting emergency care.	89
Figure 15 – Probability plot for the use of procalcitonin at different cut-offs for the diagnosis of sepsis or pneumonia in older adults visiting emergency care	91



Figure 16 – Summary ROC plot for the presence of anosmia or ageusia for the diagnosis of COVID-19 (6 studies).....	96
Figure 17 – HSROC analysis of C-reactive Protein for the diagnosis of pneumonia in adults in ambulatory care	98
Figure 18 – Diagnostic value of fever for COVID-19: bivariate meta-regression by risk of bias of participant selection (QUADAS-2).....	105

LIST OF TABLES

Table 1 – Checklist for the selection of a systematic review	18
Table 2 – Interpretation of QUADAS-2 questions on participant selection.....	27
Table 3 – Interpretation of QUADAS-2 questions on index test	31
Table 4 – Interpretation of QUADAS-2 questions on reference standard	35
Table 5 – Interpretation of QUADAS-2 questions on flow and timing.....	38
Table 6 – Comparison of QUADAS-2 and QUADAS-C	43
Table 7 – Comparison of QUADAS-2 and QUAPAS	44
Table 8 – Example: Relation of adverse event outcomes and importance for decision making	51
Table 9 – What it means: four levels of certainty of evidence	55
Table 10 – GRADE's approach for grading down the certainty of the evidence for risk of bias in a body of evidence	58
Table 11 – GRADE's approach to addressing subgroup analyses.....	62
Table 12 – Process for rating the certainty of a network estimate.....	72
Table 13 – Network meta-analysis glossary	74
Table 14 – Patient-important consequences of being classified into TP, TN, FP and FN.....	82
Table 15 – When to use which plot for presenting DTA results: overview	92
Table 16 – Summary point statistics of selected index tests from a bivariate meta-analysis.....	95
Table 17 – Comparison of sources of heterogeneity in DTA reviews.....	104
Table 18 – Strength of recommendation according to the 4 categories of the GRADE system...	110
Table 19 – Factors that influence the strength of a recommendation.....	110
Table 20 – Interpretation of strong and conditional recommendations by audience	110
Table 21 – Preference KCE for 'strong and conditional' in French and Dutch	111
Table 22 – Preference KCE for wording of recommendations in French and Dutch.....	111
Table 23 – AGREE II instrument.....	114
Table 24 – Evaluation criteria AMSTAR-2 for a systematic review	115
Table 25 – Cochrane Collaboration's risk of bias tool 1 for assessing risk of bias	123



LIST OF ABBREVIATIONS

ABBREVIATION	DEFINITION
ASI	Across-Studies Inconsistency index
CI	Confidence Interval
DI	Decision Inconsistency index
DT	Decision threshold
DTA	Diagnostic test accuracy
e.g.	Example given
EP	evidence profile
GCP	Good clinical practice
GPS	Good practice statement
GRADE	Grading of Recommendations, Assessment, Development and Evaluation
KCE	Belgian healthcare knowledge centre
MID	Minimal important difference
NMA	Network meta-analysis
OIS	Optimal information size
RCT	Randomised controlled trial
RIS	Review information size
RR	Relative risk
RRR	Relative risk reduction
SoF	Summary of Findings
VAS	Visual analogue scale

SCIENTIFIC REPORT

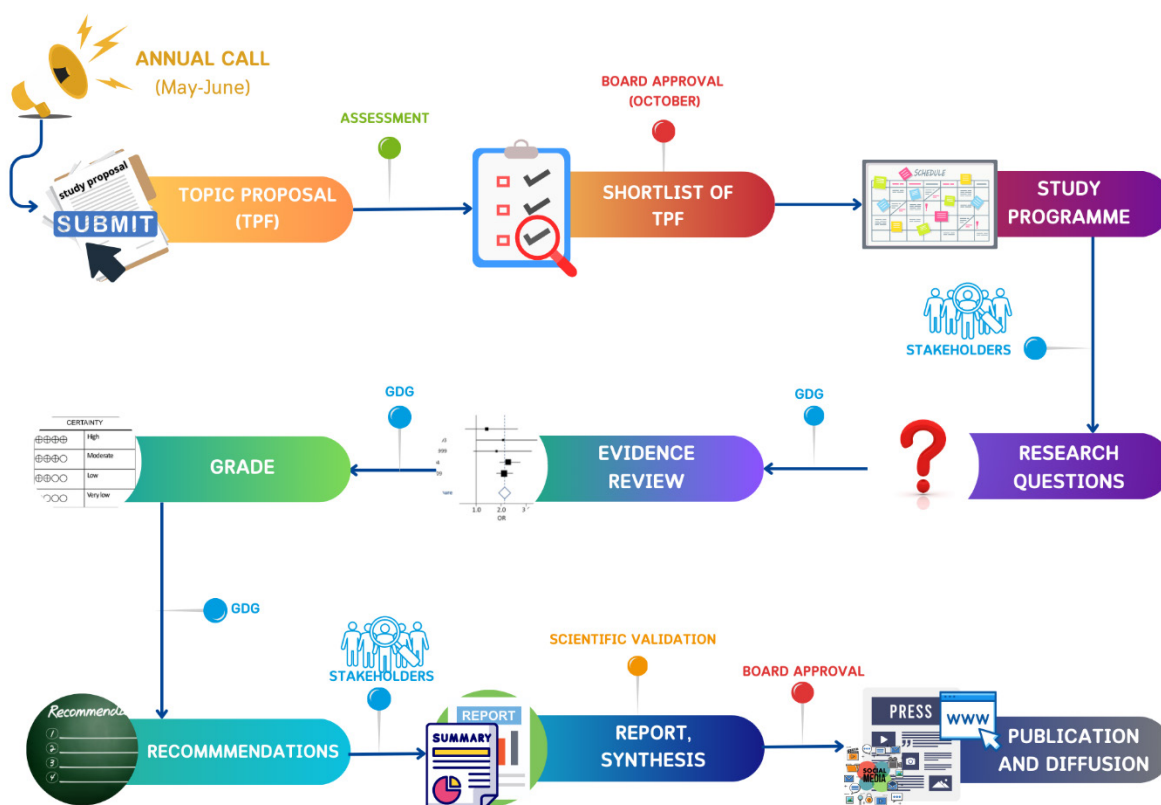
1 INTRODUCTION

1.1 Rationale

Clinical practice guidelines (named 'guideline' for the remainder of this document) are designed to improve the quality of health care and to decrease the use of unnecessary or harmful interventions. Guidelines play a crucial role in health care by providing evidence-based recommendations for diagnosing, treating, and managing various medical conditions. They help ensure consistency in care, improve patient outcomes, and reduce unnecessary variations in treatment. By synthesising the latest research, complemented with expert opinion and expert consensus, guidelines help healthcare professionals make informed decisions, balancing effectiveness with safety. They also support quality assurance, cost-effective care, and ethical medical practice.

Guidelines are developed through a rigorous and systematic process to ensure they are evidence-based and of high quality. To achieve this aim, a KCE process and methodology guide describes how to approach a new research request to develop a guideline. Figure 1 pictures the steps between the submission of new projects and the publication of KCE reports. The parts of the process typical for a guideline rely on the stringent application of methodology. For a guideline to be of high quality and evidence-based, the methodology has to be kept up to date.

Figure 1 – Standard KCE GCP report process





1.2 Aim and methods

This report aimed to update the typical GCP-methodology steps and deliver a state of the art process and methods guide for GCP projects. We set out at identifying and reviewing recent methodologies.

We applied the following methods:

- A narrative literature review including the most recent sources from the GRADE working group, Cochrane, the Latitudes network and library for validated assessment tools, Guideline International Network (GIN), the Dutch Cochrane, etc.
- An internal consultation of KCE experts with previous experiences in guideline development.

This report presents the new GCP process and methods book with the new methodology sections indicated by 'Update 2025'.

2 GENERAL ASPECTS OF GUIDELINE DEVELOPMENT

2.1 The guideline development group

A KCE guideline is developed by KCE experts in collaboration with a multidisciplinary group of health professionals named a guideline development group (GDG) with expertise in the topic studied. At least one patient representative is desirable but not mandatory. Depending on the research topic, KCE researchers may also choose to involve representatives from NIHD and health insurance organisations. An extensive search for (clinical) experts in Belgium on the topic, for example, as authors on relevant publications, or leading or recognised experts in the field is performed. Individuals or universities are contacted and asked for their interest to participate. It might be helpful to give an introduction on GRADE to the GDG to improve the working process.

The roles assigned to the GDG are:

- Definition of the clinical research questions with the KCE team (see chapter 3, prioritisation section 3.2 and development of the PICO section 3.3)
- For the review of the literature
 - Feedback on outcomes and thresholds for clinical relevance (see later chapters)
 - Identification of missed papers and clarification on the selection of papers
 - Feedback on the retrieved evidence, e.g. indirectness (applicability)
- Feedback on the content of the guideline and its recommendations (excludes the process of determining the certainty of evidence)
 - Feedback on health benefits, harms, and burden to the individual patient, certainty in the best available evidence
 - Feedback on cost, feasibility, equity from a population and patient perspective
 - Adapt content and formulation to the local context where needed

2.2 Stakeholders

Stakeholders are representatives of the relevant professional societies, patients' representatives, or other end users of the guideline. Stakeholders are individuals or groups who are affected by a guideline's development and implementation. Stakeholders are usually not involved in the guideline development group. The stakeholders provide a patient perspective on the priority of the research questions during the scoping phase. Not only on patient perspectives. Stakeholders play a key role in prioritising the research questions. For instance, they can highlight needs identified by healthcare professionals, such as the requirement for guidance on specific interventions. They review the draft

or final recommendations for clarity of the wording, feasibility and acceptability for patients, socio-cultural acceptability, feasibility for the health system, health equity, and give comments on the practical considerations to facilitate uptake and implementation of recommendations. The stakeholders' advice is independent from the GDG. Candidates that might be interested to take part as a stakeholder are sought early on; an extensive search in Belgium is performed, for example, patient organisations or professional organisations (individuals or representatives of organisations) are contacted and asked for their interest to participate.

2.3 Patient involvement

At KCE, the aim is to identify a representative mandated to speak for a patients' association involved in the topic of the guideline. If a dedicated patients' association does not exist, the umbrella patients' association of the different federal states are contacted. Again, if no candidates are identified a patient(s) will be considered on individual suitability for each project. Patient involvement in (policy) research at KCE is described in a [process note](#). This process provides guidance on how to involve patients at various stages of a research process.

2.4 Conflict of interest and KCE policy (Update 2025)

All clinicians belonging to the Guideline Development Group, as well as the stakeholders and the experts involved in the validation process complete a declaration of interest form. Information on potential conflicts of interest is published in the colophon of the report related to the guideline. All researchers of the KCE make yearly declarations of interest and details are available upon [request](#).

The GDG members are selected for their knowledge and experience, and should not represent their organisation(s) or any organisations with a direct commercial interest. Conflicts of interests can arise as different audiences are likely to have different perspectives on defining the questions. The importance of outcomes will vary when considered from the perspective of patients, clinicians or policy-makers. In this context, it should be clear for the GDG as a whole and for each GDG-member what perspective is taken, the perspective of the society as a whole or the patient perspective.

To improve the validity of the conflict of interest procedure for the GDG-members, the KCE team can share the individual submitted declarations to the entire GDG. This method worked well for the melanoma guideline.¹ It reinforces an early check of all GDG-members and will make sure all candidates can be included in the GDG. Checking for conflicts of interest among experts in a guideline development group is essential to maintaining transparency, credibility, and trust in clinical recommendations. This happens in two steps, first the following are reviewed

- Disclosures – financial relationships, research funding, industry ties, and other potential biases
- Professional Affiliations – links to pharmaceutical companies, medical device manufacturers, or organisations that could influence recommendations
- Funding Sources – Independent health organisations or government agencies may have fewer biases compared to stakeholders supported by industry?

Next, to improve on transparency, all the conflict of interests are shared with the GDG at the start of the project. For projects lasting more than a year, a renewal of the disclosure is now advised annually.



2.5 Decision-making process (Update 2025)

The GDG together with the KCE team must come to collective decisions at various stages, such as defining review questions, interpreting evidence, and formulating recommendations. While there's no fixed method for decision-making, most committees use informal consensus, and the chosen approach should always be documented.

The meetings are chaired by the KCE team which:

- Ensures inclusive participation, including lay members.
- Facilitates open, constructive debate.
- Confirms that either a majority of panellists (60% to 75%) support the final decision/recommendations or the majority (60% to 75%) does not have major objections to the proposed items.
- Documents any lack of consensus and takes the final responsibility (see colophon).

A formal consensus within the GDG is not standard practice.

- It may be used in exceptional cases (e.g., disagreement or lack of evidence).
- May involve voting or structured methods like Delphi or nominal-group methods.
- Must be transparently documented in meeting minutes and guideline methods.

An informal consensus is the more commonly used approach with, for example, the KCE researchers drafting the recommendations to be discussed at the GDG meeting after a presentation of the evidence review. An agreement on the direction of the effect, the strength, and the wording of the recommendation is then made by reaching an informal consensus. If no consensus is reached at the time of the meeting, the recommendation is adapted based on the discussion at the meeting. Next, a new proposal can be circulated via mail to the GDG or discussed at a subsequent meeting. Experts are asked to provide arguments in case of disagreement or to add clarifications where needed. After the meeting, the final recommendations (and guidance) on all research questions is again circulated inclusive of the evidence to recommendation section and experts are asked for their final comments and agreement.

The final recommendations are usually presented at a stakeholders meeting. Based on the comments received of both groups final changes are made by the KCE team.

3 DEFINING THE SCOPE AND DEVELOPING REVIEW QUESTIONS

3.1 Search for existing guidelines

In preparation of the scoping discussions, a systematic review for recent guidelines is performed and documented:

- searches of e.g. OVID Medline, the TRIP database and the G-I-N library database,
- searches of websites of scientific knowledge and professional organisations e.g.
 - Belgium: Cebam Digital Library for Health (CDLH vzw), Domus Medica
 - Europe: the National Institute for Health and Care Excellence (NICE) England, SIGN - Health Improvement Scotland, Federatie Medisch Specialisten Nederland, die Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften (AWMF) Leitlinienregister for Germany, Haute Autorité de Santé (HAS) for France.
- For guidelines in oncology, examples are
 - European Society for Medical Oncology
 - The United States: the National Cancer Institute at NIH, the American Society of Clinical Oncology (ASCO), the Society for Immunotherapy of Cancer, the U.S. Preventive Services Task Force.
 - Canada: Cancer Care Alberta, Ontario Health Care.
 - Australia: Cancer Council Australia, the Royal Australian College of General Practitioners

The identified guidelines are then appraised with the [AGREE II instrument](#).^{2,3}

3.2 Prioritisation of research questions

For defining the final scope and priority research questions experts and stakeholders are consulted. This can be through interviews, an (online) survey with stakeholders / experts, or at a meeting. The highest ranked research questions are selected based on feasibility and available resources.

Practical advice to optimise the GDG-working process when defining the questions:

- Focus on the important questions as scope needs to be limited
- Clearly explain to the GDG the difference between a critical outcome and an important outcome, as well as the implications this distinction has on the grading of the certainty of evidence.
- Explain the details for each research question with its relevant outcomes.
- Ask experts beforehand to give ratings with justifications for each outcome (e.g. excel sheet).

3.3 PICO's and outcomes

Once the scope, the aim and objective of the guideline are clear, the final clinical research questions are developed.

- The research questions are developed in collaboration with the members of the GDG.
- The Participants–Interventions–Comparator–Outcomes (PICO) framework is defined for each question. A PICO table can include the following items (depending on the type of review): Objective; Review questions; Population; Intervention; Comparator; Reference standard; Target condition; Outcomes (critical, important, cost, other); Study design; and/or Subgroup analyses.



3.4 The GRADE ADAPT and ADOPT approach (Update 2025)

3.4.1 *Adapting or adopting recommendations from a source guideline*

The GRADE- ADAPT and ADOPT approach (also called ADOLOPMENT approach) suggests to adopt or adapt recommendations from existing recent evidence-based guidelines that fulfil certain criteria for credibility and rigour of development.⁴

If a recent evidence-based guideline, developed using rigorous methods is available, adoption or adaptation of the recommendations can be considered to avoid duplication of work and save on time and resources. Figure 2 describes the process. At KCE, during the scoping phase of a guideline project, it may become clear that one or more research questions have been addressed by a recent guideline. The KCE researchers may thus decide to start from the identified source guideline of relevance and adapt or adopt certain recommendations.⁴

‘Adoption’ means the use of an existing trustworthy recommendation without modifications and without providing information on implementation for a specific situation. ‘Adaption’ means the use of an existing trustworthy recommendation with modifications such as a change in a specific element of the PICO, for example, the age of the population, or the certainty of evidence (when new evidence is identified), and may provide additional information on the local conditions for implementation.⁵

3.4.2 *Identifying a relevant source guideline of quality*

Both adoption and adaptation of recommendations starts with the identification and critical appraisal of published guidelines that could serve as source.⁴

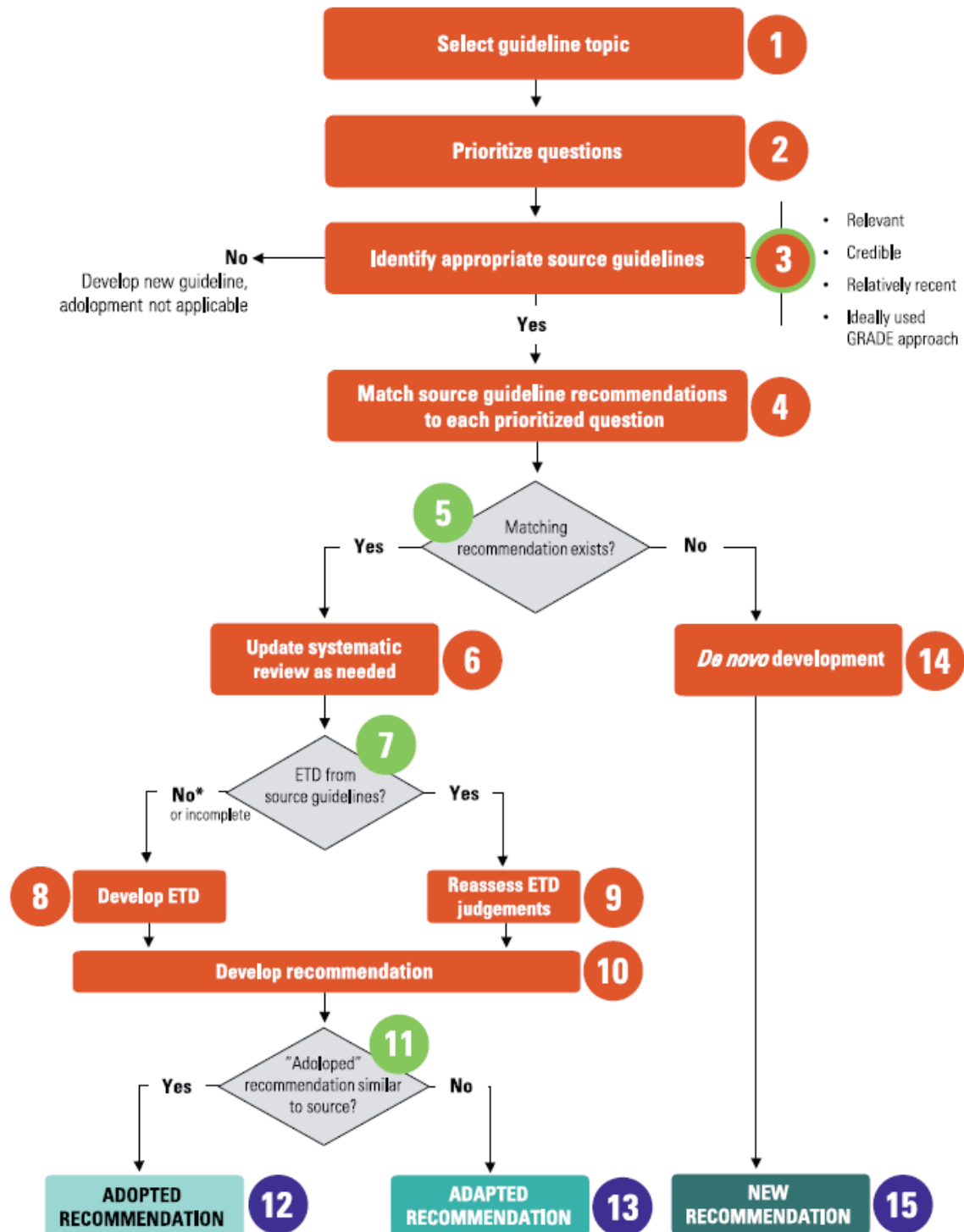
The four deciding characteristics when selecting a source guideline at KCE are relevance, credibility, recency, and the application of GRADE:

- **Relevance:** The content of the research question and the recommendation in the source guideline should match the KCE project scope. One or more research question(s) may be covered by the same guideline; similarly, a second guideline can be chosen for another research question.
- **Credibility:** the focus is on the rigour of development described in domain 3 of the [AGREE II checklist](#); Guidelines ideally need a good score for this domain (for example more than 70% but this is topic depended).⁴
- **Recency:** the source guideline is usually not older than 3 years but this can depend on the developments in the research topic under study; Attention should be paid to the date of literature searches.
- **Application of GRADE** (section 5 and 6): the guideline should present evidence tables, GRADE tables, and apply a transparent Evidence-to-Decision (EtD) framework (see section 7.1).

3.4.3 *Deciding on adapt and adopt and next steps*

The decision to adopt or adapt can be taken separately for each recommendation (Figure 2). If for a prioritised research question no credible recommendation is identified, a de novo recommendation is developed (step 5 Figure 2).

Figure 2 – The GRADE ADAPT and ADOPT approach



ETD: Evidence-to-Decision

Figure copied from Klugar 2024⁴ (GRADE: The Grading of Recommendations, Assessment, Development and Evaluations)



- *Step 1: The priority topics are identified by KCE researchers during the scoping phase on the basis of the TPF*

Example from melanoma guideline

The KCE -team searched during this scoping phase for melanoma guidelines. The AGREE II tool rigour of development was used to assess the quality of the guidelines. The most recent and evidence based guidelines were selected. From these guidelines the topics were identified. Next, a survey with experts and stakeholders was set up to prioritise the topics.

- *Step 2: Guideline research questions are prioritised, for example, based on stakeholders' priorities and agreed with GDG.*

Example

The KCE experts reviewed the priority list, drafted the list of research questions which were then scored by the GDG.

- *Step 3 with decision point: The KCE team searches for guidelines (updated search), subsequently, evaluates the retrieved guidelines for relevance, credibility (AGREE), recency, and the application of GRADE*

No guideline(s) retrieved: stop ADAPT – ADOPT approach and develop new guideline

Guideline(s) retrieved: go to step 4

Example

The guideline (scoping) search was updated and several guidelines were identified and positively evaluated.

- *Step 4: The KCE team matches each of the recommendations from the source guideline(s) to each of the research questions*
- *Decision point 5: Is a matching recommendation identified?*

Yes: it can be used for ADAPT or ADOPT

Example

The NICE guideline and Ontario melanoma guideline were chosen as source for ADAPT or ADOPT for several research questions.

No, a *de novo* research question(s) or guideline is started (step 14)

Example

Two questions were not matching and were thus developed *de novo*.

- *Step 6: Is the systematic review(s) recent enough? Does it fit the PICO of the KCE project and the Belgian context?*

No: not up to date. Perform an additional systematic review with the source PICO

No: source PICO does not address all outcomes. Perform an additional systematic review

No: context and/or values are different. Perform an additional systematic review



Continue with ADAPT approach

Yes: Continue with ADOPT approach

- *Decision point 7: Is a complete evidence to decision (EtD) framework available for each of the question(s)?*

No: develop EtD framework. Continue with ADAPT (step 8)

No: The new evidence from the updated search (step 6) is integrated in the source EtD framework. Continue with ADAPT (step 8)

Yes: assess the source EtD framework. If fine, continue ADOPT. If new aspects are added, continue with ADOPT (step 9) For example, the KCE team with GDG makes changes to the judgements, because of the need to contextualise information related to factors of local importance (e.g., resource availability)

- *Step 10: Develop a recommendation(s) with the GRADE evidence to decision (EtD) framework (see methodology in 7.1)*
- *Decision point 11: Is the recommendation adopted without changes?*

Yes: explain the rationale why it is adopted without changes. Some of the reasoning will be in the steps above (step 12)

No: For each adapted/adopted recommendation explain the rationale why. Some of the reasoning will be in the steps above. A difference in the certainty of evidence compared to the source guideline (conditional as opposed to strong recommendation) is also an ADAPT.

3.5 Search for evidence

For each research question a systematic literature review is set up by KCE experts (optionally the work can be outsourced) except in some cases when an adopt/adapt approach is used (see 3.4). An iterative approach can be followed to avoid duplication of work. For example, the evidence retrieved from a recent systematic review can be the starting point for an update. The search strategies and searches are documented in the report.

- the iterative approach starts with a search for recently published systematic review(s) (SR); then an update of the evidence synthesis is performed by searching for relevant primary studies published after the search date of the selected SR. If no SR is identified, a complete search for primary studies is performed.
- studies with a specific study design are identified e.g. SR, (diagnostic) randomised controlled trials (RCTs), observational (comparative) studies and/or prospective/retrospective diagnostic accuracy studies (in which the index test(s) and the reference standard test are applied to the same patients in a cross-sectional design). The design included is defined by the type of review (e.g. intervention vs diagnostic). The following study types are usually excluded: Narrative reviews; Case reports; Studies presented as conference abstract only or if no full text was available; and studies in which it is not possible to extract the required data.
- Databases to be used as a minimum for a search of a SR
 - Medline Ovid
 - At least one other electronic database
- Databases to be used as a minimum for a search of primary studies
 - Cochrane Library (Cochrane Reviews and Trials)
 - Medline Ovid
 - Embase
- The selection of publications and studies is performed by one (or two) researcher in two phases.
 - Phase one: screening of the titles and abstracts and excluding studies not fulfilling the inclusion criteria



- Phase two: full text screening of selected studies. If no full-text is available, the study is not taken into account. Studies published in a language other than English, Dutch, French or German are usually not included. A table with reason for exclusion for each study is presented.

3.5.1 Selection of systematic reviews (Update 2025)

The same criteria for selecting guidelines (see 3.4.2) apply here: relevance, credibility, recency, and the application of GRADE. A systematic review that would qualify as a source for a guideline is evaluated with AMSTAR-2 (see 4.2.1) and for the minimal requirements for stating that the GRADE approach was used.⁶

KCE researchers can check the AMSTAR-2 critical domains and the GRADE aspects with the following list in Table 1. **Additional aspects** important for KCE researchers are added under the KCE column. If one source needs to be selected and more than one systematic review is identified for a particular research question, priority is on rigour of methodology first and next recency (based on the literature search date). All AMSTAR-2 and GRADE criteria might not always be fulfilled. At KCE, item 4, 9, and 11 of AMSTAR-2 are considered mandatory. For a ADOPT or ADAPT procedure the items from GRADE: 'Certainty assessed for each critical or important outcome in 4 GRADE categories' and 'Evidence tables or summary of findings tables presented', will be required when the SR is considered as the basis of the guideline.

Table 1 – Checklist for the selection of a systematic review

	KCE	AMSTAR-2	GRADE
<i>Protocol registered before commencement of the review</i>		Item 2	
<i>Formulates clearly the in- and exclusion criteria</i>	X		
<i><u>Adequacy of the literature search</u></i>		Item 4	
<i>Recency of the literature search</i>	X		
<i>Addresses at least one of the research question(s)</i>	X		
<i>Evaluates at least one of the selected (critical or important) outcomes</i>	X		
<i>Searches more than one database and search is not restricted to MeSH terms</i>	X		
<i>Justification for excluding individual studies</i>		Item 7	
<i><u>Risk of bias from individual studies being included in the review</u></i>		Item 9	X
<i><u>Appropriateness of meta-analytical methods</u></i>		Item 11	

Consideration of risk of bias when interpreting the results of the review		Item 13	X
Assessment of presence and likely impact of publication bias		Item 15	X
Certainty of evidence* - thresholds or range of effect was pre-specified			X
Other GRADE domains for assessing certainty evidence -imprecision, -inconsistency, -indirectness, -large effects, -dose response gradients, -residual plausible opposing bias			X
Certainty assessed for each critical or important outcome in 4 GRADE categories			X
Evidence tables or summary of findings tables presented			X

*the GRADE definition has been used (see section 5.5): "The certainty that the true effect, accuracy measure, or association lies on one side of a particular threshold, or in a particular range." ^{7, 8} And for evidence from reviews of qualitative research: "an assessment of the extent to which the review finding is a reasonable representation of the phenomenon of interest."

3.5.2 Selection of primary studies

A selected primary study that would qualify is

- a randomised intervention trial and/or (for example if non-randomised intervention trials are identified or the RCT evidence is very limited) a non-randomised, or comparative observational intervention study (see the guidance by [Cochrane](#))⁹;
- a randomised diagnostic study or a study presenting diagnostic test accuracy (DTA) measures (for example a cross-sectional study presenting DTA) (see the guidance by [Cochrane](#))⁹;
- addresses the target patient population(s) and the comparisons in the PICO;
- evaluates at least one of the selected critical or important outcomes.



4 CRITICAL APPRAISAL OF THE LITERATURE

4.1 Introduction (Update 2025)

Critical appraisal of articles is a crucial part of a literature review. It aims at identifying methodological weaknesses and assessing the quality in a coherent way. The methodological assessment is based on a number of key questions that focus on those aspects of the study design that have a significant influence on the validity of the results reported and conclusions drawn. These key questions differ according to the study type, and a range of checklists can be used to bring a degree of consistency to the assessment process.

The process of critical appraisal consists of an evaluation by two independent reviewers who confront their results and discuss them with a third reviewer in case of disagreement. However, because of feasibility reasons, it could be acceptable that one reviewer performs the quality appraisal and that a second reviewer checks the other's work.

If necessary, the authors of the evaluated study should be contacted for additional information.

The results of the critical appraisal should be reported transparently.

Below, recommended appraisal checklists for different study types will be discussed. In addition, other commonly used – but not necessarily recommended – tools are briefly highlighted, and this for some reasons. First, when performing a systematic review of the literature, one may encounter other reviews that used a particular appraisal tool. Knowledge about this tool can help the reviewer interpret the quality assessment and conclusions in the identified review. Second, when using a published review as evidence base, and when applying the GRADE methodology to this evidence base, assessment of the risk of bias is a crucial step. A good understanding of the appraisal tool that was used can help with the decision whether or not to rate down for risk of bias.

4.2 Systematic reviews

At KCE the AMSTAR-2 is the preferred instrument and should be used.

Depending on the focus of the critical appraisal of a systematic review and the skills of the appraiser, one may choose between two commonly used and validated appraisal tools (or use them both): the AMSTAR-2 instrument or the ROBIS tool. AMSTAR-2 is considered to be more straightforward to use, while the ROBIS tool is more useful when rating reviews that included a meta-analysis.¹⁰

A third tool (ROB-ME) can be considered for assessing risk of bias due to missing evidence in systematic reviews with meta-analysis.¹¹

4.2.1 AMSTAR-2

The main aim of the Assessing the methodological quality of systematic reviews version 2 (AMSTAR-2) tool is to assess the methodological quality of a review.¹² It is made up of 16 items and has simpler response categories than the original AMSTAR version. Seven items are considered to be critical domains by the authors:

- protocol registered before commencement of the review (item 2);
- adequacy of the literature search (item 4);
- justification for excluding individual studies (item 7);
- risk of bias from individual studies being included in the review (item 9);
- appropriateness of meta-analytical methods (item 11);
- consideration of risk of bias when interpreting the results of the review (item 13);

- and, assessment of presence and likely impact of publication bias (item 15).

AMSTAR-2 can be applied to reviews of both randomised and non-randomised studies; it is primarily intended for assessing effectiveness reviews. The authors strongly recommend not to combine the individual item ratings to create an overall score.

At KCE, typically, the critical appraisal is performed by one researcher. The tool is available with this link: [AMSTAR-2](#) or via the AMSTAR-2 guidance document on the following [website](#).

4.2.2 ROBIS (Update 2025)

The main aim of the Risk of Bias in Systematic Reviews (ROBIS) tool is to evaluate the level of bias present within a systematic review.¹³

The tool is made up of three distinct phases:

- Firstly, there is an optional first phase to assess the applicability of the review to the research question of interest.
Example:

For a diagnostic review:

Category	Target question	Review being assessed ³
Patient(s):	Adults with symptoms suggestive of rectal cancer	Unclear
Index test(s):	Endoscopic ultrasound	Endoscopic ultrasound
Reference standard:	Surgical histology	Surgery
Target condition:	Early rectal cancer (T0)	Early rectal cancer (T0)

Relevance assessment	Reasoning
Does the question addressed by the review match the question you are trying to answer (e.g. in your overview or guideline)?	Unclear The question matches for the index test, reference standard and target condition categories but the population was not defined in the review and so it is unclear whether the question addressed by the review matches the target question.

Copied from [Guidance on how to use ROBIS](#)

- The second phase is made up of 20-items within four main domains: study eligibility criteria, identification and selection of studies, data collection and study appraisal, synthesis and findings. The objective of this phase is to identify concerns about the review conduct. Each domain has signalling questions and ends with a judgement of concerns of each domain (low, high or unclear).
- There is also a third phase consisting of three signalling questions to enable an overall assessment of bias rating to be given. ROBIS has a wide application and is intended for assessing reviews about effectiveness, DTA, prognosis and aetiology.

The tool was developed at the University of Bristol and next to the tool, resources (guidance, training webinar, access database, template for tabular and graphical display) to help with the application and presentation of ROBIS are available from their [website](#). The tool is available with this link: [ROBIS](#) or via the [website](#).



4.2.3 ROB-ME (Update 2025)

The Risk Of Bias due to Missing Evidence (ROB-ME) tool is designed for authors or users of systematic reviews to assess risk of bias due to missing evidence in a pairwise meta-analysis of the effect of interventions.¹¹ “Missing evidence” refers to situations in which relevant study results are absent from the meta-analysis, either because entire studies were not identified, were excluded, or because specific outcomes were selectively unreported. Such omissions can arise from publication bias, selective outcome reporting, or data availability issues. ROB-ME evaluates the extent to which these missing results could bias the meta-analysis findings and thereby influence the conclusions about the effects of the interventions. Application of the ROB-ME tool to a systematic review consists of four steps:

- Step 1, at the protocol stage: select and define which meta-analyses will be assessed for risk of bias due to missing evidence;
- Step 2, during data collection or when assessing risk of bias in the results of included studies: determine which studies meeting the inclusion criteria for the meta-analyses have missing results (e.g. outcomes that were part of the study protocol, but not mentioned in the final publication) and thus cannot contribute to the meta-analyses;
- After generating meta-analysis results, but before assessing certainty in the body of evidence:
 - Step 3: consider the potential for missing studies across the systematic review;
 - Step 4: assess risk of bias due to missing evidence in each meta-analysis.

The information gathered in the first three steps is used to answer eight signalling questions in step 4, which is completed for each meta-analysis assessed for risk of bias.

More information can be found on the following website: <https://www.riskofbias.info/welcome/rob-me-tool>.

4.3 Randomised controlled trials for interventions (Update 2025)

At KCE the RoB 1 is the preferred instrument and should be used. Teams are encouraged to pilot the tool on a subset of included studies and document signalling questions and decision rules to increase consistency across assessments.

4.3.1 RoB 1 tool

The original Cochrane Risk of Bias tool, introduced in the Cochrane Handbook (2008), evaluates seven specific domains⁹:

- Sequence generation;
- Allocation concealment;
- Blinding of participants and personnel;
- Blinding of outcome assessment;
- Incomplete outcome data;
- Selective outcome reporting;
- Other potential sources of bias.

Each domain is rated as having low, high, or unclear risk of bias, based on predefined criteria and supporting quotes from the article. Performance bias (blinding) and attrition bias (incomplete outcome data) should be assessed for each critical and important outcome separately. If insufficient detail is reported of what happened in the study, the judgement will usually be “unclear risk of bias”. Some methodological issues, such as the correctness of the statistical analysis, power, etc. are not

specifically addressed in this tool, and should be assessed separately. The Cochrane Handbook provides detailed guidance on the use of RoB 1: [website link](#).

Although RoB 1 is still widely used, it has limitations, particularly in terms of inter-rater reliability and clarity of guidance for complex trial designs.

4.3.2 RoB 2 tool

The RoB 2 tool is the updated version and represents a more structured and outcome-specific approach. It includes five domains:

- Bias arising from the randomisation process;
- Bias due to deviations from intended interventions;
- Bias due to missing outcome data;
- Bias in measurement of the outcome;
- Bias in selection of the reported result.

Each domain is assessed using a set of signalling questions with predefined algorithms to arrive at a domain-level judgement (low risk, some concerns, high risk). An overall risk of bias judgement is also made for each outcome assessed. RoB 2 requires careful planning and training but increases consistency and transparency in judgements.

The tool is available [online](#) with supporting documents and Excel templates.

Comparison of RoB 1 and RoB 2 and use in practice

Both the original Cochrane Risk of Bias tool (RoB 1) and the updated version (RoB 2) are widely used in critical appraisal of randomised controlled trials, but their application differs in terms of complexity, granularity, and required resources. RoB 2 has been developed to improve transparency and reliability in risk of bias assessments. It introduces structured signalling questions and prespecified algorithms to guide judgements within five domains, thereby minimizing subjectivity and enhancing reproducibility.¹⁴ Unlike RoB 1, which assessed most elements of the risk of bias at the study level, RoB 2 is more structurally applied at the outcome level, which allows more nuanced assessment when multiple outcomes or time points are reported. Despite these advances, RoB 2 has been shown to be resource-intensive and challenging to implement in practice. In a large-scale review of complex interventions, Crocker et al. reported a mean of 358 minutes of staff time per study to complete RoB 2 assessments, including individual reviews and consensus meetings.¹⁵ The number of outcomes, reports per study, and reviewer experience all significantly affected the time required. Even with extensive training, implementation required operational decisions not always covered by the existing guidance. Reviewers faced particular challenges in applying the guidance on:

- Deviations from intended interventions: especially when blinding was not feasible, and the potential for contamination or compensatory behaviour in control groups had to be judged without clear empirical support;
- Missing outcome data: where thresholds for acceptable missingness were not predefined in the tool, leading reviewers to adopt arbitrary cut-offs;
- Outcome measurement: especially for patient-reported outcomes, where lack of blinding was often present but not necessarily impactful;
- Selection of reported results: where the absence of a publicly available analysis plan often led to judgements of “some concerns,” even for seemingly straightforward outcomes.

These findings underscore the importance of calibration exercises and review-specific guidance when applying RoB 2, as also recommended by the tool's developers. Teams are encouraged to pilot the tool on a subset of included studies and document decision rules to increase consistency across



assessments. Moreover, the distinction between RoB judgements per outcome and overall judgements per study has practical implications for synthesis and GRADE assessments. The Cochrane Handbook recommends stratifying meta-analyses by overall risk of bias and integrating domain-level information into certainty ratings for each outcome.

In conclusion, while RoB 2 provides a conceptually robust and structured framework, its successful implementation requires methodological expertise, extensive training, and adequate time allocation. Its outcome-specific focus aligns well with modern synthesis methods such as GRADE and network meta-analysis, but review authors should anticipate and plan for the considerable workload involved. Therefore, the pragmatic choice for using the RoB 1 tool is considered acceptable.

4.3.3 ROBUST-RCT

The Risk Of Bias in Understanding the Strength of Trials—Randomised Controlled Trials (ROBUST-RCT) tool is a recently developed instrument intended to address some of the practical limitations encountered with RoB 2.¹⁶ ROBUST-RCT aims to strike a balance between methodological rigor and usability by simplifying domain structures and reducing the cognitive burden on assessors. It retains the core principle of assessing bias per outcome and incorporates structured signalling questions, like RoB 2 does.

Given its novelty and evolving status, ROBUST-RCT is not yet endorsed by major review organisations such as the Cochrane Collaboration or GRADE, and should be applied cautiously and with reference to established tools and is therefore not yet implemented at KCE.

4.3.4 Other tools

Several older or less comprehensive tools have historically been used to assess the quality or risk of bias in RCTs. While these instruments may offer ease of use or have been widely cited, they are no longer considered appropriate for formal critical appraisal in systematic reviews due to methodological limitations.

- **Jadad scale:** This early instrument assesses trials based on reporting of randomisation, blinding, and withdrawals, yielding a numerical score (0–5). Although simple and easy to apply, the Jadad scale has been criticized for its narrow focus, poor coverage of key bias domains (e.g. allocation concealment), and reliance on reporting rather than actual conduct of the study. It also encourages overall scoring, which is discouraged in contemporary methodology.
- **PEDro scale:** Commonly used in physiotherapy and rehabilitation research, the PEDro scale includes 11 items, several of which assess reporting quality rather than risk of bias. Like the Jadad scale, it produces an overall score, which is discouraged by current guidance due to the risk of misleading aggregation across unrelated domains.

4.4 Diagnostic accuracy studies (Update 2025)

At KCE the QUADAS-2 is the preferred instrument and should be used. Teams are encouraged to pilot the tool on a subset of included studies and document signalling questions and decision rules to increase consistency across assessments.

Diagnostic test accuracy studies provide essential evidence on whether tests correctly identify people with or without a target condition. However, such studies are prone to bias, and their results may not always apply to the intended clinical context.

4.4.1 Introduction to QUADAS-2

The Quality Assessment of Diagnostic Accuracy Studies (QUADAS) version 2 tool helps reviewers evaluate the risk of bias and applicability of diagnostic accuracy studies in a systematic and transparent way.

This section provides practical guidance on how to use and interpret the QUADAS-2 tool. It's intended to supplement existing handbooks and references (cited below) by offering a step-by-step approach. You'll find worked examples that illustrate how to apply the tool, as well as a standardised template designed to ensure consistency among multiple researchers working on the same review. This guidance will help you align your team, adapt the signalling questions to your specific research question, and effectively judge the risk of bias.

For a full elaboration on the development of the QUADAS-2 tool and its use, please refer to: Whiting PF et al.¹⁷, Reitsma et al.¹⁸, Davenport CF, et al.¹⁹

4.4.1.1 Overview of the QUADAS-2 Tool

QUADAS-2 assesses study quality across four domains:

- **Participant Selection** – Were participants selected appropriately?
- **Index Test** – Was the test under evaluation conducted and interpreted correctly?
- **Reference Standard** – Was the comparator test or 'gold' standard valid and applied properly?
- **Flow and Timing** – Did all participants receive the same tests in a timely manner, and were all participants included in the analysis?

For each domain, two independent reviewers judge:

- Risk of bias (Low, High, or Unclear)
- Concerns about applicability (Low, High, or Unclear)

QUADAS-2 is not a scoring system; instead, judgements are made per domain. Reviewers must tailor signalling questions to their review context!

4.4.1.2 Preparing to Use QUADAS-2

Before applying the tool:

- Define the review question clearly, using PICO for DTA (PIRTS): Participants, Index test, Reference Standard, Target disease, Setting.
- Define the population of interest clearly (e.g. adults with suspected pulmonary embolism, children undergoing screening for developmental delay).
- Specify whether the focus is on initial diagnosis, differential diagnosis, or screening.
→ This clarity guides the adaptation of signalling questions.
- Customise signalling questions to fit the index test(s), target disease, reference standard, and clinical setting under review.
- Pilot test the tool on a few studies to check consistency among team members.
- Train and calibrate reviewers, resolving differences before full assessment.



4.4.1.3 Step-by-Step Application of the QUADAS-2 tool

For efficiency and transparency, we recommend developing a standardised template (e.g. in Excel or an equivalent data-extraction tool hosted on a secure online platform). This template should be adapted to the specific review question and piloted on three to five included studies.

Piloting serves two purposes:

- To ensure consistency of judgements among reviewers.
- To verify whether the signalling questions are appropriate and sufficiently capture potential sources of bias in the context of the review.

The signalling questions should be adapted, expanded, or removed depending on the review topic. Additional items may also be added when the standard QUADAS-2 guidance does not fully capture relevant methodological issues.

4.4.1.4 QUADAS-2 common pitfalls and tips

- Avoid overusing “Unclear” – attempt to contact authors or check protocols if possible.
- Do not ignore applicability concerns – these affect generalisability.
- Ensure signalling questions are tailored to your review (e.g. different for imaging vs lab tests). QUADAS-2 is a tool, not a strict guideline. Review teams should adapt and tailor the tool to their review. This is explained for each of the domains in the next sections.
- Document decision-making transparently to improve reproducibility of the review:
 - Document all adaptations made to the QUADAS-2 questions.
 - Justify why certain items were added or removed.
 - Keep a version history of the template.
 - Store completed templates in the secure shared platform to maintain traceability.

4.4.2 Domain 1: Participant Selection

Note: We use the term "participant" selection instead of "patient" selection because not everyone included in a diagnostic study is necessarily a patient. For example, some individuals suspected of a disease may turn out to be healthy, and studies on screening often include people who are not symptomatic.

4.4.2.1 Step 1: Describing basic study characteristics

Before assessing the risk of bias for this domain, you should describe the items that are crucial for your research question. For example, you might want to consider the following items:

- Purpose/aim of the study
- Study design (prospective/retrospective cohort/case-control/...)
- Study setting (e.g. Primary, Secondary, or tertiary care, GP, ED, mixed settings, ...)
- Sample size (n with target disease and N total)
- Inclusion and exclusion criteria
- Country + year
- Demographics

Recording these items in the template will facilitate discussions on differences in bias and applicability judgement among team members. These items should be tailored to the specific review question.

4.4.2.2 Step 2: Adapt the signalling questions to your review context

The original (standard) QUADAS-2 tool provides three standard signalling questions for risk of bias in participant selection:

Q1. Was a consecutive or random sample of participants enrolled?

Q2. Was a case–control design avoided?

Q3. Did the study avoid inappropriate exclusions?

Please refer to the *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy* for an elaborate explanation on how to interpret these questions. Table 2 shows a brief overview on how to interpret these questions.

Table 2 – Interpretation of QUADAS-2 questions on participant selection

Question	Interpretation	Purpose of the question
1. Was a consecutive or random sample of participants enrolled?	Ensures that participants reflect the real-world spectrum seen in practice (e.g. clinical setting, symptom severity, prior testing). If selection is non-consecutive or non-random, bias may be introduced by systematically including certain types of participants (e.g. only severe cases).	Helps detect spectrum* bias. Strong preselection of participants can distort test accuracy estimates. Including only obvious cases inflates sensitivity, while selecting overly 'clean' controls inflates specificity.
2. Was a case–control design avoided?	Case–control designs (e.g. comparing diseased vs. healthy volunteers) often overestimate test accuracy because they tend to include extreme groups, not reflecting real screening or diagnostic practice.	Avoids selection bias inherent in artificial contrasts between diseased and healthy groups. Case–control studies often yield more optimistic results than prospective cohort studies in representative clinical populations.
3. Did the study avoid inappropriate exclusions?	Studies should not exclude participants in a way that distorts the diagnostic spectrum—e.g. excluding those with borderline test results, comorbidities, or those who couldn't complete reference testing. Excluding such participants may lead to overoptimistic accuracy estimates.	Prevents bias from systematic exclusions that skew the study sample toward "ideal" participants.

*Spectrum bias refers to the selection of patients across only a limited part of the 'severity spectrum' of disease e.g. when only severely affected patients are included in the study they will not be representative of the whole spectrum of patients.

Modify these questions so they are specific to your review. For example:

- If the review concerns emergency settings, you might rephrase Q1 as:
"Were all eligible participants presenting consecutively in the emergency department included, without selective recruitment (e.g. excluding weekend admissions)?"
- If the review concerns rare diseases, you might refine Q2:
"Was the study design free from extreme case–control sampling (e.g. only severe cases vs. healthy volunteers)?"



- For Q3, adapt based on typical exclusions in the field:
“Did the study avoid excluding participants with comorbidities, borderline test results, or those unable to complete the reference test?”

Adaptation of the questions avoids vague judgements and ensures consistency across reviewers.

4.4.2.3 Step 3: Add Review-Specific Items Where Needed (Optional)

Examples of additional signalling items that might be added:

- “Were participants recruited prospectively rather than retrospectively?”
- “Were inclusion/exclusion criteria defined prior to recruitment?”
- “Were participants representative of the setting in which the test would be applied in practice (e.g. primary care vs. tertiary referral centre)?”

Consider adding these items to capture subtle biases that the standard framework might not cover.

4.4.2.4 Step 4: Pilot the questions and your template on 3–5 studies

- Assign two or more reviewers to independently apply the adapted questions.
- Compare responses for agreement.
- Identify ambiguities (e.g. reviewers interpreting “inappropriate exclusions” differently, see example below).
- Revise the template until consistent judgements can be made.
- Example of different interpretation by two reviewers, concerning inappropriate exclusions:

Example Rapid Antigen Tests for COVID-19; Study design of a hypothetical study:

A study of a rapid antigen test for SARS-CoV-2 excluded all participants who:

- Had symptoms for more than 5 days, or
- Were unable to provide a nasopharyngeal swab (e.g. frail elderly participants, children), or
- Had inconclusive PCR results (the reference standard).

Reviewer A’s interpretation: High risk of bias

- These exclusions remove precisely the types of participants in whom the test is most likely to perform less well (mild/late disease, young/old participants, technically difficult sampling).
- Excluding these groups creates an ‘ideal’ population. This typically leads to overestimation of sensitivity, and sometimes specificity.
- Therefore, Reviewer A scores this as “inappropriate exclusions” → high risk of bias.

Reviewer B’s interpretation: Low risk of bias

- The 5-day symptom cutoff reflects the test’s regulatory indication (the manufacturer only recommends use in early symptomatic cases).
- Excluding participants unable to undergo swabbing is reasonable because the test cannot be applied in them anyway.
- Excluding inconclusive PCR results avoids misclassification bias in the reference standard.
- Reviewer B therefore considers these exclusions appropriate → low risk of bias.

Why disagreement arises:

- The difference hinges on how narrowly the review question is defined:
- If the review asks, “How well do rapid antigen tests work in all participants presenting with possible COVID-19?” → Reviewer A is right.
- If the review is restricted to “participants with ≤5 days of symptoms, for whom a nasopharyngeal swab can be performed” → Reviewer B is right.

“Inappropriate exclusions” often depend on whether the excluded participants are relevant to the review’s target population. Reviewers need to agree on the population definition up front, otherwise judgements will diverge!

Remark on independent reviewing: The Cochrane Handbook recommends that the QUADAS-2 tool should be applied by at least two people, independently. While the handbook doesn’t explicitly mandate blinding to the other reviewer’s judgements during the assessment, the requirement for an independent assessment serves the same purpose. The core principle is to avoid one reviewer’s opinion from influencing another’s judgement before a consensus is reached. The review team must decide on the best way to ensure independence. Blinding is a more rigorous method, but a non-blinded approach might be more pragmatic.

4.4.2.5 Step 5: Record Judgements Systematically

In your Excel (or online) template:

- For each signalling question, provide Yes/No/Unclear options.
- Optionally, reviewers can be asked to substantiate their assessment with citations or references to the study (e.g. citing page/line in article, or stating “unclear, methods not described”).
- Summarise the risk of bias judgement (Low / High / Unclear) for the participant selection domain, by answering the final question: “Could selection of participants have introduced bias? Low risk, unclear risk, high risk”.

How to assign each category:

- **Low risk:**
If *all* relevant signalling questions are answered “Yes”, and
There are no additional concerns about participant selection beyond those covered by the signalling questions.
- **High risk:**
If *one or more* signalling questions are answered “No”, and this could have introduced bias in relation to the review question.
Example: a case–control design (Q2 = “No”) always leads to high risk.
- **Unclear risk:**
If one or more signalling questions are “Unclear” *and* the available information is insufficient to judge whether selection introduced bias.
Example: A study doesn’t describe how participants were recruited, and it’s not possible to infer from the context.
(Remark: consider contacting the authors)

In summary

Low risk = All “Yes” (or an occasional “Unclear” that clearly does not threaten validity).

High risk = Any “No” that could reasonably distort test accuracy for the review question.

Unclear risk = *Insufficient detail* to judge (e.g. recruitment method not reported).



4.4.2.6 Applicability

Summarise the **applicability** judgement (high concern, unclear concern, low concern) for the participant selection domain, by answering the final question: “Is there concern that the included participants do not match the review question?”

Look for:

- Demographics: Do the ages, sexes, or other characteristics of the study participants align with the target population of the review?
- Clinical context: Do the participants' levels of disease severity, symptom presentation (e.g. onset of symptoms), or previous testing align with the clinical scenario described in the review's target question?
- Exclusions: Were participants excluded in a way that might make the study results less applicable to the intended target population?
- Specific populations: If the review is focused on a specific subgroup (e.g. children, participants with a certain symptom score), do the study participants represent this subgroup?

Example:

A review assessing the accuracy of a rapid strep test in children with a sore throat might find a study has high concern for applicability if it included adults or a disproportionate number of children without symptoms suggestive of strep throat, or if it excluded children who would typically be tested in the target setting.

How to avoid common mistakes:

- Do not assume participant recruitment was consecutive just because it was prospective. Example: An ultrasound study says “participants were prospectively included” but omits whether they were consecutive. Risk should be *unclear*, *not* automatically *low*. However, if the study states, “We enrolled *all* eligible participants prospectively,” you can likely assume that consecutive enrolment was used.
 - When participant enrolment methods aren't clearly stated, the review team must determine how to handle this ambiguity. While the safest approach is to contact the authors for clarification, this can be a time-consuming process, and authors may not always respond.
- Do not underestimate the impact of case–control designs. By design, case–control studies almost always inflate accuracy (spectrum bias), and QUADAS-2 guidance is clear: this should always be rated as high risk of bias of bias. Example: A test for Alzheimer's disease using participants with confirmed dementia vs. healthy volunteers will overestimate both sensitivity and specificity.
- Distinguish bias versus applicability issues. Bias is about study conduct; applicability is about the fit to your own review question. Example: A study of mammography in women aged 40–49 only may be less applicable to a review covering all adults, but the study itself could still be low risk of bias if sampling was appropriate.
- Do not give benefit of the doubt when information is missing. QUADAS-2 guidance recommends using “unclear” if reporting does not permit judgement. Example: A study doesn't describe how participants were chosen from hospital records → should be denoted as unclear risk, not low or high. Less experienced researchers tend to perceive missing information as “low risk”, while more experienced researchers often take the opposite view.
- Do not overlook subtle exclusions, ignoring exclusions that seem “practical” but actually distort the spectrum. Excluding participants with comorbidities, indeterminate results, or who could not undergo the reference standard often inflates accuracy. Example: An ultrasound study excludes

obese participants “because of image quality issues.” This removes exactly the type of participants where performance might be worse.

4.4.3 Domain 2: Index test

4.4.3.1 Step 1: Describing index test items

Before assessing the risk of bias for this domain, you should describe the items that are crucial for your research question. For example, in a review about the signs and symptoms of a specific infection, you might want to consider the following:

- The name of the sign or symptom: What is/are the specific sign(s) or symptom(s) being evaluated?
- The test method: How was the sign or symptom detected or measured?
- Timing of the test: When was the test performed relative to the onset of the infection?
- Who applied the test: Who administered the test or made the observation?
- Definition of a positive result: What criteria were used to define a positive outcome?
- Blinding to the reference standard: Was the person interpreting the test result aware of the final diagnosis?
- Predetermined threshold: Was the cutoff for a positive result established before the study began?

Again, these items should be tailored to the specific review question.

4.4.3.2 Step 2: Adapt the signalling questions to your review context

The original (standard) QUADAS-2 tool provides two standard signalling questions for risk of bias in the index test:

Q1. Were the index test results interpreted without knowledge of the results of the reference standard?

Q2. If a threshold was used, was it pre-specified?

Please refer to the *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy* for an explanation on how to interpret these questions. Table 3 shows a brief overview on how to interpret these questions.

Table 3 – Interpretation of QUADAS-2 questions on index test

Question	Interpretation	Purpose of the question
1. Were the index test results interpreted without knowledge of the results of the reference standard?	Checks whether the person interpreting the index test was blinded (i.e., unaware) of the reference standard results.	Ensures blinding, thus avoiding reviewer's bias—interpreting the index test with knowledge of the reference outcome could skew the index test interpretation. This is particularly important for subjective tests like imaging or clinical judgement. It may sometimes be less important for lab tests.
2. If a threshold was used, was it pre-specified?	Determines whether the cut-off or decision threshold for a positive/negative test was defined a priori, before seeing study data.	Prevents data-driven threshold selection, which can lead to optimistic (inflated) accuracy estimates; pre-specification supports validity and reproducibility. Selecting a threshold post hoc (after reviewing results) tends to overestimate diagnostic performance.



Modify these questions so they are specific to your review. For example:

- For a review of MRI for multiple sclerosis (MS), you might refine Q1: “Were the MRI scans interpreted without knowledge of the participant’s clinical diagnosis or reference standard findings (neurological exam and follow-up)?”
- For prostate-specific antigen (PSA) for prostate cancer, you might refine Q2: “Was the PSA threshold (e.g. 4 ng/mL) defined in advance, or was it selected based on study results?”
- For blood pressure measurement devices, you might refine Q2: “Was the diagnostic threshold for hypertension (e.g. $\geq 140/90$ mmHg) defined according to accepted guidelines before study initiation?”

4.4.3.3 Step 3: Add Review-Specific Items Where Needed (Optional)

Examples of additional signalling items that might be added:

- “*Were the index test interpreters trained and calibrated before performing/reading the test?*” Such a question might be important for studies on imaging tests (e.g. MRI, ultrasound), or clinical examination. Assesses bias from poorly trained or inconsistent test readers/assessors.
- “*For tests requiring subjective grading (e.g. Alvarado score for suspected appendicitis, imaging scores): was a standardised scoring system used?*” Assesses variability in interpretation of tests/assessments.

Consider adding these items to capture subtle biases that the standard framework might not cover.

4.4.3.4 Step 4: Pilot the questions and your template on 3–5 studies

As for all domains:

- Assign two or more reviewers to independently apply the adapted questions.
- Compare responses for agreement.
- Identify ambiguities
- Revise the template until consistent judgements can be made.

Example of different interpretation by two reviewers, concerning index test

Hypothetical Study: Use of the Alvarado Score for diagnosing appendicitis in adults

Reference standard: Histopathology after appendectomy.

The study states: “All participants underwent scoring at admission, and the final diagnosis was based on histology.

The score was calculated retrospectively from case notes.”

The paper does not mention whether clinicians calculating the score were blinded to histology. The paper also reports test accuracy at different cut-offs (≥ 5 , ≥ 7 , ≥ 8), but does not say whether these were pre-specified.

Reviewer A’s interpretation: High risk of bias for the overall Index Test domain

Signalling question 1 (Blinding): “Were index test results interpreted without knowledge of reference standard results?”

Says “No” (high risk of bias), because the score was calculated retrospectively and blinding is not mentioned. Reviewer assumes clinicians likely knew the diagnosis from case notes or surgery outcome.

Signalling question 2 (Threshold pre-specified):

Says “No” (high risk of bias), because multiple thresholds were reported without clear pre-specification.

Reviewer B’s interpretation: Unclear risk of bias for the overall Index Test domain

Signalling question 1 (Blinding):

Says “Unclear”, because the paper doesn’t explicitly state whether assessors were blinded. Reviewer gives benefit of the doubt and avoids assuming bias.

Signalling question 2 (Threshold pre-specified):

Says “Unclear”, because although multiple thresholds were presented, the authors don’t clarify whether these were pre-set or exploratory.

Conclusion: Reviewer A interprets the lack of reporting as likely bias, while reviewer B interprets the lack of reporting as uncertainty. This kind of difference is common and usually resolved through discussion or consulting a third reviewer. The review team should agree on how to handle these situations.

4.4.3.5 Step 5: Record Judgements Systematically

In your Excel (or online) template:

- For each signalling question, provide Yes/No/Unclear options.
- Optionally, reviewers can be asked to substantiate their assessment with citations or references to the study (e.g. citing page/line in article, or stating “unclear, methods not described”).
- Summarise the **risk of bias** judgement (Low / High / Unclear) for the index test domain, by answering the final question: “Could the conduct or interpretation of the index test have introduced bias? Low risk, unclear risk, high risk”.

How to assign each category:

- **Low risk:**
If all relevant signalling questions are answered “Yes”, and
There are no additional concerns about the index test beyond those covered by the signalling questions.
- **High risk:**
If *one or more signalling questions are answered “No”*, and this could have introduced bias in relation to the review question.
- **Unclear risk:**
If one or more signalling questions are “Unclear” *and* the available information is insufficient to judge whether bias was introduced.
(Remark: consider contacting the authors)

In summary

Low risk = All “Yes” (or an occasional “Unclear” that clearly does not threaten validity).

High risk = Any “No” that could reasonably distort test accuracy for the review question.

Unclear risk = *Insufficient detail* to judge (e.g. test method not reported).

4.4.3.6 Applicability

Summarise the **applicability** judgement (high concern, unclear concern, low concern) for the index test domain, by answering the final question: “Are there concerns that the index test, its conduct, or interpretation differ from the review question?”

When judging applicability for the index test, ask yourself:

- Is this the same test as in my review question?
 - Was it performed the way it would be in real practice?
 - Was it interpreted in the same way, by similar people, in a similar context?
- If the answer is “no” (or “not sure”), then there are concerns about applicability.



How to avoid common mistakes:

- Do not assume blinding: if not stated, mark “Unclear” (unless the study protocol or supplementary materials contain more information).
- Look carefully for thresholds: pre-specification is crucial! Multiple exploratory cut-offs usually means high risk of bias.
- Distinguish bias versus applicability issues. Bias is about study conduct; applicability is about the fit to the review question.
- Discuss reviewer differences in calibration meetings — align on when to rate “high” vs “unclear.” Do this early on in the review process.

4.4.4 Domain 3: Reference Standard

4.4.4.1 Step 1: Describing the reference standard

Before assessing the risk of bias for this domain, you should describe the items that are crucial for your research question. For example, you might want to consider the following items:

- What was the reference standard for cases, including the threshold (if applicable)
- Which samples were used (if applicable)
- What was the timing of reference standard, when was it conducted?
- Was it interpreted with blinding to the index test?
- Did it incorporate the index test?

These items should be tailored to the specific review question.

4.4.4.2 Step 2: Adapt the signalling questions to your review context

The original (standard) QUADAS-2 tool provides two standard signalling questions for risk of bias in the reference standard:

Q1. Is the reference standard likely to correctly classify the target condition?

Q2. Were the reference standard results interpreted without knowledge of the results of the index test?

Please refer to the *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy* for an explanation on how to interpret these questions. Table 4 shows a brief overview on how to interpret these questions.

Table 4 – Interpretation of QUADAS-2 questions on reference standard

Question	Interpretation	Purpose of the question
1. Is the reference standard likely to correctly classify the target condition?	Examines whether the chosen reference standard is valid, accurate, and appropriate for defining the “true” disease status. A poor or imperfect reference standard can lead to misclassification.	Ensures the reference standard is credible. If it cannot correctly classify the target condition (e.g. clinical follow-up instead of histology for cancer), then the index test’s accuracy will be biased.
2. Were the reference standard results interpreted without knowledge of the results of the index test?	Checks whether the person interpreting or applying the reference standard was blinded to the index test results. If not blinded, knowledge of the index test may influence their judgement.	Prevents review bias (also called diagnostic review bias). Blinding ensures that classification by the reference standard is independent of index test results.

Modify these questions so they are specific to your review. For example:

- For a COVID-19 review (the loss of sense of smell as index test, PCR as reference), you might refine Q2: “Were the PCR results interpreted without knowing whether participants had a loss of sense of smell?”
- For a breast cancer review (mammography as index test, biopsy as reference), you might refine Q2: “Were the histopathology results interpreted without knowledge of the mammogram findings?”

Note that some reference standards are less subjective than others. This will impact the final judgement for this domain. For example, a lab result will be less prone to bias from knowledge of the index test results than a reference test that requires subjective interpretation.

4.4.4.3 Step 3: Add Review-Specific Items Where Needed (Optional)

Examples of additional signalling items that might be added:

- “If more than one reference standard was used, were they applied consistently across participants?” Such a question might be important for reviews where some participants get histology, others get follow-up or imaging (common in appendicitis, oncology, paediatrics). Assesses differential verification bias, as different reference standards may classify disease differently.
- “Does the reference standard incorporate index test results?” Assesses whether index test and reference standard are independent of each other. If not, this usually inflates the apparent accuracy of the index test, because a positive index test can help define a “positive disease” classification. Consider adding this question when composite reference standards are used.

4.4.4.4 Step 4: Pilot the questions and your template on 3–5 studies

As for all domains:

- Assign two or more reviewers to independently apply the adapted questions.
- Compare responses for agreement.
- Identify ambiguities
- Revise the template until consistent judgements can be made.



Example of different interpretation by two reviewers, concerning the reference standard

Hypothetical study on appendicitis diagnosis:

Reference standard: Histopathology of the removed appendix after surgery

Not all participants got the same reference standard: for some participants who were managed conservatively, the final diagnosis was based on clinical follow-up

Blinding of pathologists was not reported

Reviewer A's interpretation: High risk of bias for the overall reference standard domain

Signalling question 1: No (mixed standards = inconsistent classification).

Signalling question 2: No (assumes pathologists might know clinical information).

Overall: High risk of bias.

Reviewer B's interpretation: Unclear risk of bias for the overall reference standard domain

Signalling question 1: Unclear (histopathology is good, but worried about those without surgery).

Signalling question 2: Unclear (not reported; but in practice, pathologists are usually blinded).

Overall: Unclear risk of bias.

Conclusion: Reviewer A tends to interpret lack of reporting as evidence of bias (conservative, stricter), while reviewer B tends to interpret lack of reporting as uncertainty (less strict, gives benefit of the doubt). These differences are usually resolved by discussion or consulting a third reviewer, or sometimes after checking trial protocols or contacting the study authors.

4.4.4.5 Step 5: Record Judgements Systematically

In your Excel (or online) template:

- For each signalling question, provide Yes/No/Unclear options.
- Optionally, reviewers can be asked to substantiate their assessment with citations or references to the study (e.g. citing page/line in article, or stating "unclear, methods not described").
- Summarise the **risk of bias** judgement (Low / High / Unclear) for the reference standard domain, by answering the final question: "Could the reference standard, its conduct, or its interpretation have introduced bias? Low risk, unclear risk, high risk".

How to assign each category:

- **Low risk:**
If *all* relevant signalling questions are answered "Yes", and
There are no additional concerns about the reference standard beyond those covered by the signalling questions.
- **High risk:**
If *one or more* signalling questions are answered "No", and this could have introduced bias in relation to the review question.
- **Unclear risk:**
If one or more signalling questions are "Unclear" *and* the available information is insufficient to judge whether bias was introduced.
(Remark: consider contacting the authors)

In summary

Low risk = All "Yes" (or an occasional "Unclear" that clearly does not threaten validity).

High risk = Any "No" that could reasonably distort test accuracy for the review question.

Unclear risk = *Insufficient detail* to judge.

4.4.4.6 Applicability

Summarise the **applicability** judgement (high concern, unclear concern, low concern) for the reference standard domain, by answering the final question: “Are there concerns that the reference standard, its conduct, or interpretation differ from the review question?”

When judging applicability for the reference standard, ask yourself:

- Does the study use the same reference standard your review is focused on?
- Was the reference standard applied in a way consistent with your review question?
- Was the reference standard interpreted as it would be in practice, or differently?
- Does the reference standard apply equally across the population of interest?
- Was the reference standard a composite or modified version that differs from what your review intends?

If the answer is “no” (or “not sure”), then there are concerns about applicability.

How to avoid common mistakes:

- Distinguish bias from applicability issues. Bias is about study conduct; applicability is about the fit to the review question.
- Do not assume blinding: if not stated, mark “Unclear” (unless the study protocol or supplementary materials contain more information).
- Watch out for partial or composite reference standards.

4.4.5 Domain 4: Flow and Timing

4.4.5.1 Step 1: Describing flow and timing of the study

The Flow and Timing domain is the fourth and final domain of the QUADAS-2 tool. Before assessing the risk of bias for this domain, you should describe the items that are crucial for your research question. For example, you might want to consider the following items:

- Was there an appropriate time interval between index test and reference standard?
- Did all participants receive the same reference standard?
- Were all participants included in the analysis?

These items should be tailored to the specific review question.

4.4.5.2 Step 2: Adapt the signalling questions to your review context

The original (standard) QUADAS-2 tool provides three standard signalling questions for risk of bias in the Flow and Timing domain:

Q1. Was there an appropriate interval between index test(s) and reference standard?

Q2. Did all participants receive the same reference standard?

Q3. Were all participants included in the analysis?

Please refer to the *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy* for an explanation on how to interpret these questions. Table 5 shows a brief overview on how to interpret these questions.

**Table 5 – Interpretation of QUADAS-2 questions on flow and timing**

Question	Interpretation	Purpose of the question
1. Was there an appropriate interval between index test(s) and reference standard?	Assesses whether the time gap between the index test and the reference standard was short enough that the disease status would not plausibly change. Long intervals risk misclassification if the condition progresses, resolves, or fluctuates.	Ensuring both tests (index test and reference standard) evaluate the same clinical state. It depends on the type of target disease (e.g. chronic versus acute) what can be seen as an 'appropriate' time interval.
2. Did all participants receive the same reference standard?	Evaluates whether every participant was verified with the same reference method. Using different reference standards across subgroups may classify disease differently and distort accuracy estimates.	To assess the potential for differential verification bias caused by multiple reference standards.
3. Were all participants included in the analysis?	Checks whether all recruited participants were analysed, including those with indeterminate, missing, or uninterpretable results. Excluding participants selectively may inflate or deflate accuracy.	To assess the potential for selection bias and ensure results reflect the full study sample.

Modify these questions so they are specific to your review. For example:

- For a review on appendicitis diagnosis (index test: Alvarado score, reference standard: histopathology), you might refine Q1: "Was the interval between applying the Alvarado score and performing appendectomy with histopathology short enough to ensure appendicitis status did not change?"
- For a breast cancer review (mammography as index test, biopsy as reference standard), you might refine Q2: "Did all participants with suspicious findings undergo tissue biopsy as the reference standard?"
- For a review on an AI radiology tool for pneumonia diagnosis (index test = AI algorithm, reference standard: clinician diagnosis), you might refine Q3: "Were all participants included in the statistical analysis, including those with uninterpretable chest images or missing clinical data?"

4.4.5.3 Step 3: Add Review-Specific Items Where Needed (Optional)

Examples of additional signalling items that might be added:

- "Was follow-up complete for all participants?" Such a question might be important for reviews where diagnosis relies on follow-up rather than an immediate reference standard (e.g. dementia, cancer recurrence, appendicitis managed non-surgically).
- "Was there consistency in timing across participants?" Such a question might be important for reviews on acute conditions (myocardial infarction, infection, trauma) where timing can affect accuracy.

4.4.5.4 Step 4: Pilot the questions and your template on 3–5 studies

As for all domains:

- Assign two or more reviewers to independently apply the adapted questions.
- Compare responses for agreement.
- Identify ambiguities
- Revise the template until consistent judgements can be made.

Example of different interpretation by two reviewers, concerning Flow and Timing

Hypothetical study on appendicitis diagnosis:

Population: Participants with suspected appendicitis in the emergency department.

Index test: Alvarado score (clinical scoring system).

Reference standard: Histopathology after appendectomy (for operated participants), or clinical follow-up for 7 days (for conservatively managed participants).

Timing: Surgery performed within 48 hours of index test. Follow-up not completed for 10% of conservatively managed participants.

Participant flow: 15% of enrolled participants excluded from analysis due to “insufficient data.”

Reviewer A’s interpretation overall: high risk of bias

Signalling question 1: Yes, for surgical cases (short interval); No, for conservative cases (7-day follow-up is too short to exclude delayed appendicitis). → high risk

Signalling question 2: No: Different standards were used (histopathology vs clinical follow-up). → high risk

Signalling question 3: No: 15% excluded, plus incomplete follow-up for some. → high risk

Reviewer B’s interpretation overall: unclear risk of bias

Signalling question 1: Yes, 48 hours for surgery and 7 days for follow-up are reasonable. → low risk

Signalling question 2: Mixed reference standards, but both are reasonable (histology = gold standard; short follow-up sufficient for those without surgery). → unclear risk

Signalling question 3: Exclusions were noted but not clearly systematic. → unclear risk

Conclusion: This example illustrates how differences in handling of poor reporting can lead to divergent judgements, which are usually resolved through discussion with a third reviewer or with clinical experts to clarify clinical interpretation e.g. how to define whether 7 days follow-up is sufficient or not?

4.4.5.5 Step 5: Record Judgements Systematically

In your Excel (or online) template:

- For each signalling question, provide Yes/No/Unclear options.
- Optionally, reviewers can be asked to substantiate their assessment with citations or references to the study (e.g. citing page/line in article, or stating “unclear, methods not described”).
- Summarise the **risk of bias** judgement (Low / High / Unclear) for the flow and timing domain, by answering the final question: “Could the participant flow have introduced bias? Low risk, unclear risk, high risk”.

How to assign each category:

- **Low risk:**

If the interval was appropriate, all participants received the same reference standard, and all were included in analysis.

- **High risk:**

If there were long/variable intervals, use of different reference standards, or a selective exclusion of participants.



- Unclear risk:
If reporting was insufficient to judge.
(Remark: consider contacting the authors)

In summary

Low risk = All “Yes” (or an occasional “Unclear” that clearly does not threaten validity).

High risk = Any “No” that could reasonably distort test accuracy for the review question.

Unclear risk = *Insufficient detail* to judge.

For the Flow and Timing domain, there is *no* applicability question in the original QUADAS-2 tool. Flow and Timing is about study conduct (intervals, verification, exclusions) rather than definitions of the target condition or the clinical setting.

How to avoid common mistakes:

- Distinguish bias versus applicability issues. Bias is about study conduct; applicability is about the fit to the review question. Treating problems in flow (e.g. long delays, missing data) as “applicability concerns” is not correct.
- Consider disease progression when judging timing. Do not assume that *any* interval is acceptable without considering the disease course. Think about the natural history of the condition. For acute diseases (e.g. appendicitis, MI, infection), short intervals are essential. For chronic diseases (e.g. dementia, cancer), longer intervals may be acceptable.
- Look for consistency of reference standards. Check whether *all participants* received the same reference standard. If not, assess whether this could classify disease differently and bias results.
- Make sure all participants are accounted for in the analysis (check the tables and the flow diagrams!). Ensure that all participants enrolled were included in the analysis, or exclusions were transparently justified.

4.4.6 Visualising and presenting QUADAS-2 results

4.4.6.1 Tabular presentations for Individual studies

Each study assessed is listed in rows, with QUADAS-2 domains (risk of bias + applicability concerns) in columns. The risk judgements (“low”, “high”, “unclear”) are filled in for each cell. This provides maximum transparency but can be too long if many studies are included.

Example: QUADAS-2 Risk of Bias and Applicability Concerns (Individual Studies)

Study ID	Participant Selection	Index Test	Reference Standard	Flow & Timing	Applicability Concerns (Participants / Index / Reference)
Smith 2020	High	Low	Low	Unclear	High / Low / Low
Jones 2019	Low	Unclear	Low	Low	Low / Low / Low
Struyf 2018	Unclear	Low	Unclear	High	Low / High / Unclear

4.4.6.2 Tabular presentations for Summary of studies

Instead of showing every study, summarise counts or percentages. These tables are often paired with a figure for clarity.

Example: QUADAS-2 summary of Risk of Bias across studies

Domain	Low Risk (n, %)	High Risk (n, %)	Unclear (n, %)
Participant Selection	6 (40%)	7 (47%)	2 (13%)
Index Test	9 (60%)	3 (20%)	3 (20%)
Reference Standard	11 (73%)	2 (13%)	2 (13%)
Flow & Timing	5 (33%)	6 (40%)	4 (27%)

Example: QUADAS-2 Summary of Applicability Concerns

Domain	Low Concern (n, %)	High Concern (n, %)	Unclear (n, %)
Participant Selection	12 (80%)	2 (13%)	1 (7%)
Index Test	11 (73%)	3 (20%)	1 (7%)
Reference Standard	13 (87%)	1 (7%)	1 (7%)

4.4.6.3 Graphical presentations

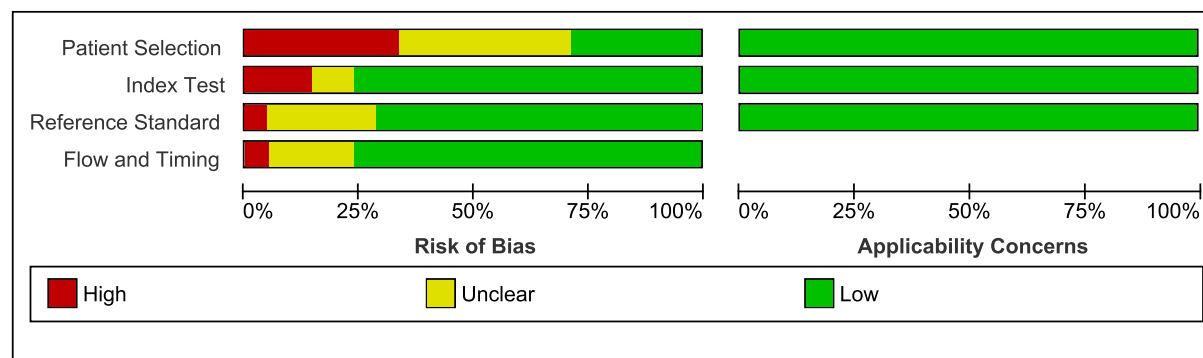
Graphical presentations (like traffic-light plots and bar charts) are often preferred in QUADAS-2 reporting, because tables with dozens of studies and multiple domains can be hard to digest. Graphs allow readers to see patterns and problem areas at a glance (e.g. "participant selection is often high risk"). A traffic-light plot immediately communicates both individual study-level risks and overall distribution, and stacked bar charts summarise proportions across domains without requiring detailed reading.

Cochrane and many journals now recommend or require graphical presentation of QUADAS-2 results, often alongside tables. While tables ensure transparency (every judgement is visible), graphs offer a synthesis, showing patterns that are easy to understand. The best practice is to include both: a detailed table (usually as supplementary material) and a figure (in the main text). Detailed study-level QUADAS-2 forms can be included in a supplementary file for transparency.

BAR CHARTS

Show the proportion of studies rated as "low", "high", or "unclear" risk of bias per domain. These stacked bar plots summarise across all studies for each QUADAS-2 domain. They are useful for a quick overview of methodological strengths and weaknesses.

Example of a bar chart, generated with Review Manager²⁰





In this example, there are no applicability concerns for this DTA review. The extent of applicability concerns typically depends on the inclusion and exclusion criteria of the review: the more strictly these criteria are defined and applied when selecting studies, the fewer applicability concerns will arise.

TRAFFIC LIGHT PLOT (PER STUDY)

Each study is listed, and coloured symbols (green+ = low, yellow? = unclear, red- = high) represent the risk judgements for each domain. This helps readers visually grasp the risk profile of each included study.

Example of a traffic light plot, generated with Review Manager²⁰

	Risk of Bias				Applicability Concerns		
	Patient Selection	Index Test	Reference Standard	Flow and Timing	Patient Selection	Index Test	Reference Standard
Camm 2018	+	+	-	-	+	+	+
Caterino 2004	-	-	?	+	?	+	+
de Gadiana Romualdo 2017	+	-	+	+	+	+	+
de Gadiana Romualdo 2018	+	-	+	+	+	+	+
Geier 2013	+	+	-	+	+	+	+
Heining 2016	+	+	?	+	+	+	+
van Nieuwkoop 2010	+	-	+	+	+	+	+

- High	? Unclear	+ Low
---	---	--

This QUADAS-2 traffic light plot can also be generated in R, using the 'robvis' package. Robvis is specifically designed for visualising risk-of-bias assessments, including QUADAS-2. It can produce both traffic-light plots (study-level, green/yellow/red) and summary bar plots (proportions across domains).

4.4.7 QUADAS-C

QUADAS-C is a tool developed to assess the risk of bias in comparative diagnostic accuracy studies, i.e. studies that directly compare two or more index tests in the same population. It is essentially an adaptation of QUADAS-2.

Both tools are structured around the same four key domains: participant selection, index test(s), reference standard, and flow and timing. QUADAS-C also uses signalling questions to guide judgements, and provides judgements of risk of bias within each domain, supporting transparent, structured assessment.

QUADAS-C focuses specifically on comparisons between tests, so the signalling questions and guidance are adapted to address comparative bias (e.g., whether both tests were applied to the same participants under the same conditions, whether interpretation was blinded equally). It adds emphasis on whether differences in patient flow, test application, or interpretation could introduce bias favouring one test over another, rather than bias in absolute test accuracy.

QUADAS-C does not substantially alter the domains themselves, but reframes them to capture risks relevant to head-to-head test evaluation. Table 6 shows a comparison of QUADAS-2 and QUADAS-C.

Table 6 – Comparison of QUADAS-2 and QUADAS-C

Domain	QUADAS-2 (single test accuracy)	QUADAS-C (comparative accuracy)
<i>Participant selection</i>	Were participants selected appropriately for evaluating accuracy of one index test?	Were participants selected in a way that allows fair comparison between the tests (e.g. same spectrum, no selective application)?
<i>Index test(s)</i>	Was the index test conducted and interpreted without bias, and blinded to reference standard?	Were <i>all index tests</i> applied and interpreted in the same participants, in the same way, and under comparable conditions (to avoid systematic advantage for one test)?
<i>Reference standard</i>	Was the reference standard likely to correctly classify the target condition, and interpreted independently of the index test?	Same as QUADAS-2, but also considers whether the reference standard might differentially advantage one test over another.
<i>Flow and Timing</i>	Did all participants receive the same reference standard, with minimal delay, and were all included in analysis?	Did all participants receive <i>all</i> tests under comparison and the reference standard, with consistent timing and analysis across groups, to prevent differential verification or attrition bias?

For a more elaborate discussion of the QUADAS-C tool, please refer to:

- Yang B et al.²¹
- Kaul T et al.²²

4.4.8 QUAPAS

QUAPAS (Quality Assessment of Prognostic Accuracy Studies) is a methodological tool developed to critically appraise the risk of bias and applicability in prognostic accuracy studies. This tool evaluates how well a prognostic marker, model, or test predicts future clinical outcomes. QUAPAS was introduced by Lee et al. (2022) as an adaptation of the QUADAS-2 tool.

Both tools use a domain-based approach to assess risk of bias and concerns about applicability. They both rely on signalling questions to guide reviewers' judgements, and both are meant to support systematic reviews by ensuring transparent and standardised quality assessment.

However, they are also very different in scope and content, as QUAPAS targets prognostic accuracy (predicting future outcomes over time). Because of this shift, QUAPAS had to modify domains and signalling questions to account for longitudinal elements such as follow-up, time-to-event outcomes, and handling of censoring. For example, QUAPAS emphasises whether outcome assessment is complete and unbiased *over time*, while QUADAS-2 emphasises blinding of reference standard and correct patient flow *at baseline*. Table 7 shows a comparison of QUADAS-2 and QUAPAS.

**Table 7 – Comparison of QUADAS-2 and QUAPAS**

Domains	QUADAS-2 (Diagnostic Accuracy)	QUAPAS (Prognostic Accuracy)
<i>Purpose</i>	Assess risk of bias & applicability in diagnostic accuracy studies (disease <i>present/absent</i> at one time point).	Assess risk of bias & applicability in prognostic accuracy studies (predicting <i>future outcomes</i> over time).
<i>Structure</i>	4 domains: (1) Participant selection, (2) Index test, (3) Reference standard, (4) Flow and timing.	4 analogous domains: (1) Participant selection, (2) Index test/prognostic model, (3) Outcome, (4) Analysis and follow-up.
<i>Focus of Index Test</i>	Diagnostic test: ability to distinguish between diseased vs. non-diseased now.	Prognostic factor/model: ability to predict who will experience an outcome in the future.
<i>Reference/Outcome</i>	Requires a reference standard.	Requires clearly defined clinical outcome(s) with appropriate follow-up.
<i>Timing</i>	Concerned with whether the index test and reference standard are measured close together in time.	Concerned with whether follow-up is long enough, complete, and unbiased.
<i>Analysis Issues</i>	Mainly patient flow, missing data, and interpretation of reference standard.	Includes handling of censoring, competing risks, appropriate statistical modelling of time-to-event data.
<i>Key Similarity</i>	Domain-based structure, signalling questions, judgements of risk of bias and applicability.	Same structure and process of signalling questions, adapted to prognosis context.
<i>Key Difference</i>	Cross-sectional, static evaluation of test accuracy.	Longitudinal, predictive evaluation that accounts for time, prognosis, and outcome ascertainment.

For a more elaborate discussion of the QUAPAS tool, please refer to:

- Lee J et al.²³
- Kaul T et al.²²

4.5 Observational studies (Update 2025)

At KCE the ROBINS-I is the preferred instrument and should be used. The Newcastle Ottawa Scale can be considered a pragmatic choice to reduce workload as an alternative to the ROBINS-I tool, provided that the individual items are reported, and the total score is not considered.

For non-randomised studies, the risk of bias can arise from a broader range of methodological issues than in randomised trials, particularly due to confounding and selection mechanisms. The appropriate appraisal tool depends on the type of non-randomised design, the aim of the study (intervention versus exposure), and the context of the review. Several organisations provide structured tools and guidance for assessing observational studies:

- The Joanna Briggs Institute (JBI) offers critical appraisal tools for a wide range of designs, including cohort, case-control, cross-sectional and prevalence studies;
- The National Institute for Health and Care Excellence (NICE) and the Scottish Intercollegiate Guidelines Network (SIGN) have their own tools or checklists for non-randomised studies used in guidelines.

4.5.1 ROBINS-I

The Risk Of Bias In Non-randomised Studies of Interventions (ROBINS-I) tool was developed by members of the Cochrane Bias Methods Group and the Cochrane Non-Randomised Studies Methods Group to assess the internal validity of non-randomised studies of the effects of interventions (NRSI).²⁴ It enables reviewers to evaluate bias in a way conceptually aligned with the RoB 2 instrument for RCTs, which facilitates integration of both study types into evidence syntheses and GRADE assessments.

Evaluations of the risk of bias of NRSI are facilitated by considering each NRSI as an attempt to mimic a “target” trial, i.e. a hypothetical pragmatic randomised trial, conducted on the same participant group and without features putting it at risk of bias, whose results would answer the question addressed by the NRSI.

ROBINS-I is structured around seven bias domains:

- Pre-intervention: (1) bias due to confounding; (2) bias in selection of participants;
- At intervention: (3) bias in classification of interventions;
- Post-intervention: (4) bias due to deviations from intended interventions, (5) missing data, (6) measurement of outcomes, and (7) selection of reported results.

Each domain includes signalling questions, which guide reviewers toward judgements of low, moderate, serious, or critical risk of bias. An overall judgement for each outcome is based on the highest risk rating across domains.

ROBINS-I is methodologically robust but often challenging to implement in practice. It requires deep knowledge of the subject matter, careful specification of the “target trial”, and a high degree of methodological expertise. Reviewers should carefully document decision rules and calibrate their assessments across team members.

Guidance and templates can be found on the following [website](#).

Only recently, the same research group launched version 2 of the ROBINS-I tool. This new version aims to simplify terminology, refine the signalling questions, and offer algorithm-based support like that in the RoB 2 instrument. A template can be found [here](#).

4.5.2 ROBINS-E

The ROBINS-E tool (Risk Of Bias In Non-randomised Studies - Exposures) extends the ROBINS framework to observational epidemiological studies (cohort studies) that examine the effects of non-manipulable exposures, such as environmental pollutants, dietary patterns, or socio-behavioural characteristics.²⁵

ROBINS-E follows the same domain structure (7 domains of bias) and logic as ROBINS-I but adapts the signalling questions to address specific challenges of exposure studies. The 7 domains include:

- Bias due to confounding;
- Bias arising from measurement of the exposure;
- Bias in selection of participants into the study;
- Bias due to post-exposure interventions;
- Bias due to missing data;
- Bias in measurement of the outcome; and
- Bias in selection of the reported result.



The current version addresses only follow-up (cohort) studies, and work is ongoing to develop ROBINS-E variants for other study designs including case-control studies.

Guidance and templates can be found on the following [website](#).

4.5.3 *Newcastle-Ottawa Scale*

The Newcastle-Ottawa Scale is a checklist originally designed for cohort and case-control studies. It covers three domains (selection, comparability, and outcome/exposure) but lacks detailed signalling questions or standard algorithms. It does not distinguish between bias and reporting quality, and allows for numerical scoring, which is no longer recommended in modern methodologies such as GRADE. It can be considered as an alternative to the ROBINS-I tool, provided that the individual items are reported and the total score is not taken into account.

4.6 Risk of bias in network meta-analysis

To assess potential biases in the conduct, analysis, and interpretation of systematic reviews applying a network meta-analysis, the Risk of Bias in Network Meta-Analysis (RoB NMA) tool can be used.²⁶ The tool should be applied separately and after the risk of bias critical appraisal with for example, AMSTAR-2 or ROBIS. For the steps and methods of this tool, we refer to the paper by Lunny et al.²⁶

5 GRADE FOR INTERVENTIONS

5.1 Introduction

5.1.1 *What is GRADE? (Update 2025)*

GRADE (Grading of Recommendations, Assessment, Development and Evaluation) offers a system for rating certainty (previously also called quality) of evidence in systematic reviews and guidelines and grading strength of recommendations in guidelines. The core system is designed for reviews and guidelines that examine alternative management strategies or interventions, which may include no intervention or current best management. Further GRADE methods focus on other research questions such as diagnosis, prognosis, and qualitative evidence; It tries to offer a transparent and structured process for developing and presenting evidence summaries for systematic reviews and guidelines in health care and for carrying out the steps involved in developing recommendations.

The GRADE approach is based on a sequential assessment of the certainty of evidence and subsequent decision about the strength of a recommendation. Separating the judgements regarding the certainty of evidence from judgements about the strength of recommendations is a critical and defining feature of the GRADE system. Therefore, the GRADE system allows that conditional recommendations in the face of high certainty evidence are common because factors other than the certainty of evidence will influence the strength of a recommendation, such as balance between desirable and undesirable effects, patient values and preferences, use of resources, feasibility, equity and acceptability. For the same reason, it allows for strong recommendations based on evidence from well-designed observational studies or when only low or very low certainty of evidence is available.²⁷ But, these are rather exceptional when the recommendation is in favour and more common when the recommendation is against.

Example of a weak recommendation based on high certainty evidence

Several RCTs compared the use of combination chemotherapy and radiotherapy versus radiotherapy alone in unresectable, locally advanced non-small cell lung cancer (stage IIIA). The overall certainty rating for overall survival was considered high; compared with radiotherapy alone, the combination of chemotherapy and radiotherapy reduced the risk for death corresponding to a mean gain in life expectancy of a few months. The overall certainty rating for the outcome harm and burden related to chemotherapy, which was increased, was also high. Taking into account the negative view of patients on the increased burden of harms (most patients would place a low value on the small survival benefit), the

small survival benefit was insufficient to make a strong recommendation and the guideline panel offered a weak recommendation despite the high certainty of the available evidence.

Example of a strong recommendation based on very low certainty evidence

The use of [¹⁸F]FDG PET/CT for the primary staging of nodular metastasis in patients with melanoma was studied in a combination of prospective and retrospective studies. The overall certainty of the accuracy outcome on sensitivity was very low. The experts considered FDG-PET/CT not useful in the evaluation of regional metastases. The burden of the test is considerable with anxiety and psychological distress due to the high number of patients incorrectly classified as having a nodular metastasis and needing further investigations. The final recommendations by KCE was strongly not to recommend ultrasound for N staging.

Although the GRADE system makes judgements about certainty of evidence and strength of recommendations in a systematic and transparent manner, it does not eliminate disagreements in interpreting evidence nor the inevitable need for judgements in deciding on the best among alternative courses of action. This is a very important aspect, and the developers of GRADE repeatedly warn against a too mechanistic approach and stress the fact that different experts can use GRADE and come to different conclusions. What matters in the first place is to ensure a maximum transparency about judgements and to explicitly acknowledge the values and preferences and other arguments that underlie the recommendation.

GRADE should not be seen as a tool for standardisation in the first place, but as a framework for quality assurance and transparency.

The “Core GRADE” papers are published in the [BMJ \(Research methods & reporting\)](#) in 2025, giving a good overview of the basic principles of GRADE. Core GRADE aims at providing a concise and structured overview of the key elements for using GRADE. It focusses on PICO that compare two interventions. Core GRADE is an excellent starting point for new users or for users who would like to update their knowledge of GRADE.²⁸

The GRADE Handbook can be found online too, which is considered the most comprehensive and up-to-date resource on the GRADE approach. The new version will be completed by 2026 here: <https://book.grade.pro.org/>.

Other guidance documents that explain how to apply GRADE can be found in the journal of Clinical Epidemiology, under the “Special issues and article collections”.

There are three collections related to GRADE:

- [GRADE guidance](#)
- [GRADE Concept Articles](#)
- [GRADE Notes](#)

Details about the specific methodology used, for example how thresholds are defined, need to be reported in each KCE report.

5.1.2 Applicability of GRADE (Update2025)

GRADE is designed for the evaluation of interventions in the first place, such as drugs, surgical interventions or radiation therapy. The GRADE methodology is continuously evolving, with recent guidance now also covering diagnostics, prognostic factors, values and preferences, network meta-analysis etc.

Some recommendations, while potentially useful, do not require formal grading of the certainty of evidence (see 7.3). These are typically recommendations for which it is sufficiently obvious that desirable effects outweigh undesirable effects to the extent that no direct evidence is available, because no one would reasonably conduct a study addressing the implicit clinical question. Typically,



such recommendations are supported by substantial indirect evidence, although identifying and detailing this evidence would be difficult and unnecessarily time-consuming. A common way to recognise these questions is that, if the alternative were made explicit, it would appear bizarre or even absurd.²⁹

5.1.3 Steps in the GRADE process

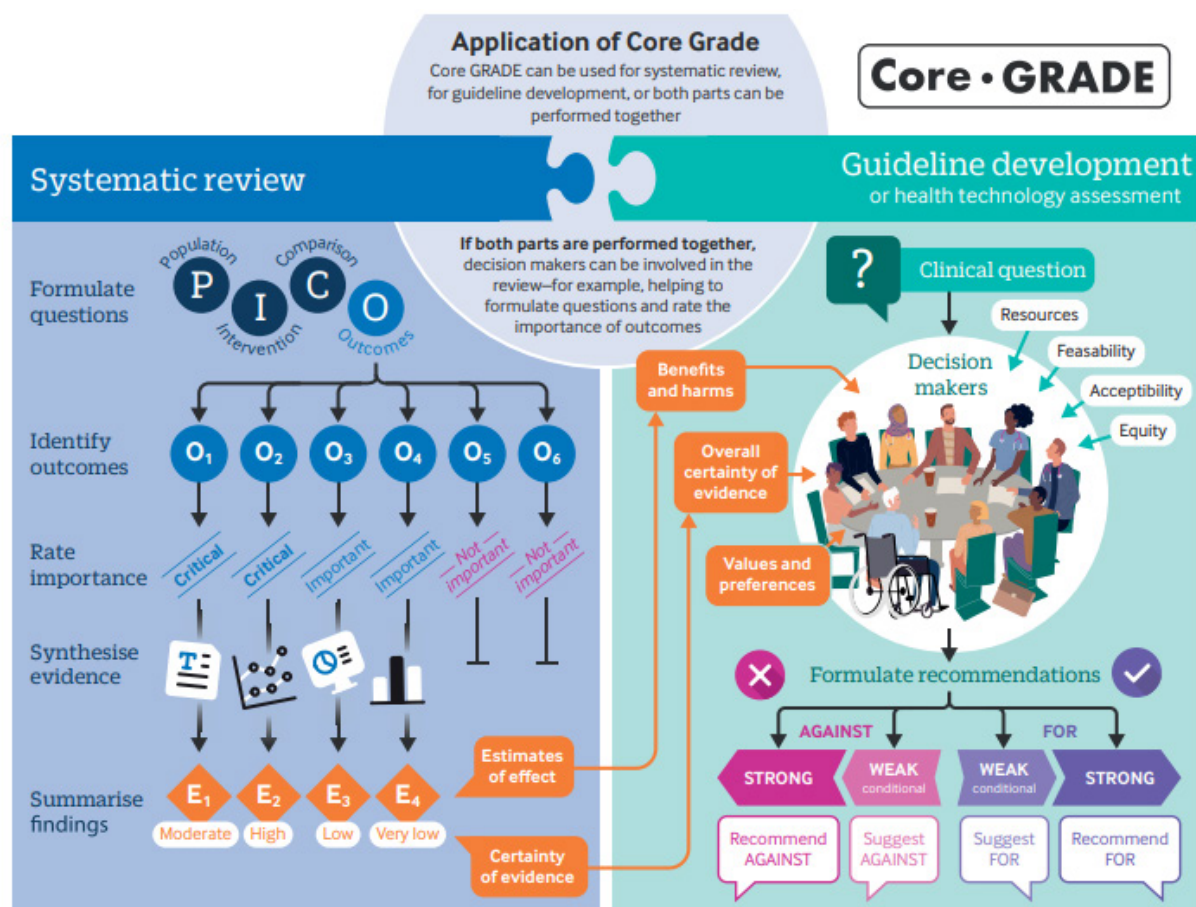
The [GRADEpro](#) software and the [MAGICapp](#) facilitates the steps in the GRADE process e.g. developing both Evidence Profiles (EP) and Summary of Findings (SoF) tables. The GRADEpro software and the MAGICapp are not further discussed and are not included in this update.

GRADE includes the following steps:

- Ask a specific healthcare question to be answered by a recommendation;
- Identify all important outcomes for this healthcare question;
- Judge the relative importance of outcomes;
- Summarise all relevant evidence;
- Grade the certainty of evidence for each outcome and for each comparison;
- Decide on the overall certainty of evidence across outcomes;
- Include judgements about the underlying values and preferences related to the management options and outcomes;
- Decide on the balance of desirable and undesirable effects;
- Decide on the balance of net benefits and cost;
- Grade the direction and strength of the recommendation;
- Formulate a recommendation;

The steps are visualised in the following Figure ³⁰

Figure 3 – Overview of the Core GRADE approach

Source Guyatt 2025 Core GRADE 1³⁰



5.2 Framing the question

GRADE requires a clear specification of the relevant setting, population, intervention, comparator(s), and outcomes with the PICO format. A guideline question often involves an additional specification, i.e. the setting in which the guideline will be implemented. For instance, guidelines intended for resource-rich environments will often be inapplicable in resource-poor environments. Furthermore, in some cases it may be necessary to specify if the guideline needs to be implemented in an inpatient or an outpatient setting.

Questions must be sufficiently specific: across the range of populations, interventions, and outcomes, a more or less similar effect must be plausible. Because the relative risk associated with an intervention vs. a specific comparator is usually similar across a wide variety of baseline risks, it is usually appropriate for systematic reviews to generate single pooled estimates of relative effects across a wide range of patient subgroups. Recommendations, however, may differ across subgroups of patients at different baseline risk of an outcome, despite there being a single relative risk that applies to all of them. Absolute risk reductions are often greater in higher-risk patients, warranting taking a higher risk of side effects and enduring inconvenience. Evidence certainty may also differ across subgroups, and this may mandate differing recommendations (higher likelihood of recommending an intervention, or making a stronger recommendation when evidence is of higher certainty). Thus, guideline panels may consider defining separate questions (and produce separate evidence summaries) for subgroups of patients.

Example

Atrial fibrillation, an irregular and often very rapid heart rhythm, can lead to blood clots in the heart and result in a cerebrovascular attack (stroke). The risk varies with age, increasing substantially for older individuals. Other factors, such as hypertension, diabetes, and a history of prior stroke, also elevate this risk. Blood thinners (anticoagulants) can prevent blood clots from forming, reducing the risk of stroke. The main side effect of these medications is an increased risk of bleeding. Warfarin requires regular blood tests and dietary monitoring. The direct Oral Anticoagulants have fewer interactions and no need for routine monitoring. As a result, a guideline panel might consider warfarin therapy for the patients at higher risk of stroke as opposed to patients with low risk of stroke. Therefore results should be presented as separate subgroups.

5.3 Deciding on the importance of outcomes

5.3.1 General approach

Outcomes may include survival (mortality), clinical events (e.g. stroke or myocardial infarction), patient-reported outcomes (e.g. specific symptoms, quality of life), adverse events, burdens (e.g. demands on caregivers, frequency of tests, restrictions on lifestyle) and economic outcomes (e.g. cost and resource use). It is critical to identify both outcomes related to adverse effects/harm as well as outcomes related to effectiveness.

GRADE specifies three categories of outcomes according to their importance. Guideline developers must, and authors of systematic reviews are strongly encouraged to specify all potential patient-important outcomes as the first step in their endeavour. The categories are:

- Critical;
- Important, but not critical;
- Of limited importance.

The first two classes of outcomes will bear on guideline recommendations; the third may or may not. Ranking outcomes by their relative importance can help to focus attention on those outcomes that are considered most important, and help to resolve or clarify disagreements. GRADE recommends to focus on a maximum of 7 critical and/or important outcomes.

Guideline developers should first consider whether particular desirable or undesirable consequences of a therapy are important to the decision regarding the optimal management strategy, or whether they are of limited importance. If the guideline panel thinks that a particular outcome is important, then it should consider whether the outcome is critical to the decision, or only important, but not critical. To facilitate ranking of outcomes according to their importance guideline developers as well as authors of systematic reviews may choose to rate outcomes numerically on a 1 to 9 scale (7 to 9 – critical; 4 to 6 – important; 1 to 3 – of limited importance) to distinguish between importance categories. Table 8 shows an example for some outcomes that have been scaled in the context of adverse events from a medical treatment.

Review authors should consider how outcomes can differ across the stages of an illness.³⁰ And if the measurement used will influence the outcome, both in terms of the type of scale used and/or the time frame of the measurement (days, years etc.).³⁰

Table 8 – Example: Relation of adverse event outcomes and importance for decision making

Importance of outcomes	Scale	Outcome*
Critical for decision making	9	Death, severe allergic reaction
	8	Heart failure
	7	Fractures
Important for decision making	6	Mood swings
	5	
	4	Headaches
Low importance for decision making	3	Nausea
	2	
	1	Itchiness

*any combination is possible

5.3.2 Perspective of outcomes

Different audiences are likely to have different perspectives on the importance of outcomes and will be different considering the perspective of patients, clinicians or policymakers.

Core GRADE suggests to perform a systematic review of patient values and preferences, and, if resources permit, a focus group of relevant patients.³⁰ At KCE, it is unusual to have the time and resources to perform this type of systematic review. On the contrary patients are involved throughout the project. The [KCE process note](#) 'Patient involvement in policy research at KCE' describes how to involve patients at various stages of a research process. Aspects that are applied addressing including the patient perspective are, a participative approach with patient involvement and the method of selecting participants aiming for a diverse patient group (from a theoretical point of view).

5.3.3 Before and after literature review

For a guideline, an initial rating of the importance of outcomes should precede the review of the evidence, and this rating should be confirmed or revised following the evidence review.

One should aim to decide which outcomes are important during protocol development and before one undertakes a systematic review or guideline project.

Guideline panels should be aware that in some instances the importance of an outcome may only become known after the protocol is written, the evidence is reviewed, or the analyses are carried out. For example, if differences in critical and important outcomes between two treatments are minimal, differences in outcomes that were initially not considered important may drive the direction and strength of a recommendation and thus become a critical outcome.³¹ The panel should take appropriate action to include these newly identified outcomes in the evidence tables.³¹



5.3.4 Implications of the classification

Only outcomes considered critical (rated 7—9) or important (rated 4—6) should be included in the evidence profile (see 5.4).

Only outcomes considered critical (rated 7—9) are the primary factors influencing a recommendation and should be used to determine the overall certainty of evidence supporting this recommendation.

When determining which outcomes are critical, it is important to bear in mind that absence of evidence on a critical outcome normally leads to a downgrading of the evidence. However, one need not to rate down confidence because of lower confidence in critical outcomes that support the recommendation which is at the same time based on critical outcomes with a high(er) confidence rating.³¹

5.3.5 Use of surrogates

Guideline developers should consider surrogate outcomes only when high certainty evidence regarding important outcomes is lacking.

Example

For melanoma treatment, the trials studying some of the new adjuvant treatments presented results of recurrence free survival only, as the follow-up time was not yet long enough for results on overall survival to be available. The recurrence free survival in this instance can be considered as a surrogate for the critical outcome of mortality (survival).

When such evidence is lacking, guideline developers may be tempted to list the surrogates as their measures of outcome. This is not the approach GRADE recommends. Rather, they should specify the important outcomes and the associated surrogates they must use as substitutes. The necessity to substitute with the surrogate may ultimately lead to rating down the certainty of the evidence because of indirectness.

5.3.6 Decision threshold and minimal important difference (Update 2025)

When decisions are made about outcomes, one should consider the minimal clinical importance of an effect size, as this is best decided before the evidence is collected and summarized, in order to avoid subjective and ad hoc decisions influenced by the available evidence.

GRADE uses the term (Clinical) Decision Threshold to refer to reference values that guide judgments about both the size of the effect (e.g., trivial, small, moderate, or large) and the associated certainty of evidence. Eventually, they also inform judgments of the balance of effects between alternative options (e.g., deciding whether benefits outweigh harms) or support commitment to an alternative intervention (e.g., recommending one intervention over another).³²

For binary outcomes this usually implies a risk reduction. The threshold is likely to differ according to the outcome, e.g. a mortality reduction of 10 % will be more important than a reduction of 10% in the number of patients developing a rash. For continuous outcomes, the minimally important difference is used, i.e. the smallest difference in outcome of interest that informed patients or proxies perceive to be important, either beneficial or harmful, and that would lead the patient or clinician to consider a change in management.

Notes

- *A minimally important difference is measured at the individual level.*
- *The effect on a continuous outcome can be expressed as a mean difference, but also as the proportion of patients having a benefit that is above the minimally important difference.*

Example

The minimally important difference (defined by the GDG) for a reduction in neck pain is a reduction of 10 on the VAS scale (visual analogue scale, a scale used to measure pain, from 1 to 100). An intervention that reduces neck pain can cause a mean reduction of 30 on the VAS scale, but it can also be expressed as a proportion, e.g. “the intervention caused a reduction of pain of more than 10% on the VAS scale” or “the minimally important difference was reached in 40% of patients”.

Determining these thresholds is not straightforward and often difficult. Expert opinion is often essential.

For a few outcomes validated thresholds exist based on evidence from surveys amongst patients, e.g. the Cochrane back pain group determined a threshold for back and neck pain. Doing a specific literature search on this topic is probably too labour-intensive and moreover, there are no universally accepted and agreed validated methods for doing so. Some rules of thumb are provided by the GRADE working group, such as an increase/decrease of 25%, but one should be cautious to apply these without a critical reflection on the context.

Initially, GRADE focussed on two possible thresholds for rating the certainty: (1) no difference between groups (e.g. RR of 1.0 or risk difference of 0), or (2) an important effect (i.e. minimally important difference, also called small effect threshold). More recent GRADE guidance offers to use a partially or fully contextualized approach, where reviewers would rate their certainty in relation to a range of effects (i.e. a trivial, small, moderate or large effect) that is bounded by two thresholds.³³ Although the terminology of minimally, partially and fully contextualised have been abandoned again, for guideline development groups, using a range of trivial, small, moderate and large effects can facilitate decision making when developing recommendations.³⁴ When developing a recommendation regarding an intervention with severe side effects such as bone marrow transplantation for example, guideline developers can choose to rate the certainty of the evidence that there is a large beneficial effect that is large enough to balance the side effects, rather than just an effect that is larger than the minimal clinically important difference. A large effect is then the target of the certainty rating.

5.3.7 Outcome ‘as yet undiscovered toxicity’ for new drugs

One might consider “as-yet-undiscovered toxicity” as an important (adverse) outcome of any new drug. Indeed, an intervention may be associated with adverse effects that are not apparent in initial studies. Thus Such toxicity becomes critical only when sufficient evidence of its existence emerges.

Example

When COX-2 inhibitors were initially marketed, myocardial infarction was one of the adverse effects of a long list of speculative possibly associated adverse effects (e.g. autoimmune syndromes, bone marrow suppression, renal failure). When an increased rate of myocardial infarction with COX-2 inhibitors emerged and became evident, the outcome was formally labelled critical.

The tricky part of this judgement is how frequently the adverse event must occur and how plausible the association with the intervention must be before it becomes a critical outcome. For instance, an observational study found a previously unsuspected association between sulfonylurea use and cancer-related mortality. Should cancer deaths now be an important, or even a critical, endpoint when considering sulfonylurea use in patients with type 2 diabetes? As is repeatedly the case, we cannot offer hard and fast rules for these judgements.

5.4 GRADE Evidence Profiles and Summary of Findings tables

Evidence from optimally conducted systematic reviews must be summarised for each patient-important outcome. For each comparison of alternative management strategies, all outcomes should be presented together. It is likely that all studies relevant to a healthcare question will not provide evidence regarding every outcome. The GRADE working group has developed specific approaches to present the certainty of the available evidence, the judgements that bear on the certainty rating, and the effects of alternative management strategies on the outcomes of interest: the GRADE evidence



profile (EP) and the Summary of Findings (SoF) table. An evidence profile is more detailed than a summary of findings table. The main difference is that it provides the detailed judgements and reasons for grading up or down per category. As such, an evidence profile provides the information for other guideline developers and validators that allows them to understand how the judgement about the level of evidence was reached.

The SoF table consists of 7 elements

- A list of critical and important outcomes, both desirable and undesirable;
- A measure of the typical burden of these outcomes (e.g. control group, estimated risk);
- A measure of the risk in the intervention group or, alternatively or additionally, a measure of the difference between the risks with and without intervention;
- The relative magnitude of effect;
- Numbers of participants and studies addressing these outcomes;
- A rating of the overall certainty in effect estimates for each outcome (which may vary by outcome);
- Comments.

On top of that GRADE provides a number of supplementary recommendations:

- A SoF table should present the seven (or fewer) most important outcomes. These outcomes must always be patient-important outcomes and never be surrogates, although surrogates can be used to estimate effects on patient-important outcomes;
- A SoF table should present the highest certainty evidence;
- When certainty of two bodies of evidence (e.g. randomised trials and observational studies) is similar, a SoF table may include summaries from both.

5.5 Rating the certainty of evidence (Update 2025)

To apply GRADE and judge the evidence soundly, it is essential that at least one person involved in developing the guideline has profound knowledge of GRADE and is more than able to implement it.

5.5.1 Introduction (Update 2025)

Formally assessing the trustworthiness of the available evidence represents a key component of the GRADE approach. In the past, the certainty of evidence was defined by GRADE as “the extent of our confidence that the estimates of the effect are correct or are adequate to support a particular decision or recommendation”. This definition suggests that rating the confidence was focussing on the point estimates of the effect. More recently, GRADE adopted a new definition that is grounded in the idea that, when deciding whether evidence regarding effects is adequate to support a recommendation, we are not assessing our confidence in point estimates of effects, but rather our confidence in where these effects lie relative to particular thresholds.³⁵

GRADE's adopted definition of certainty of the evidence

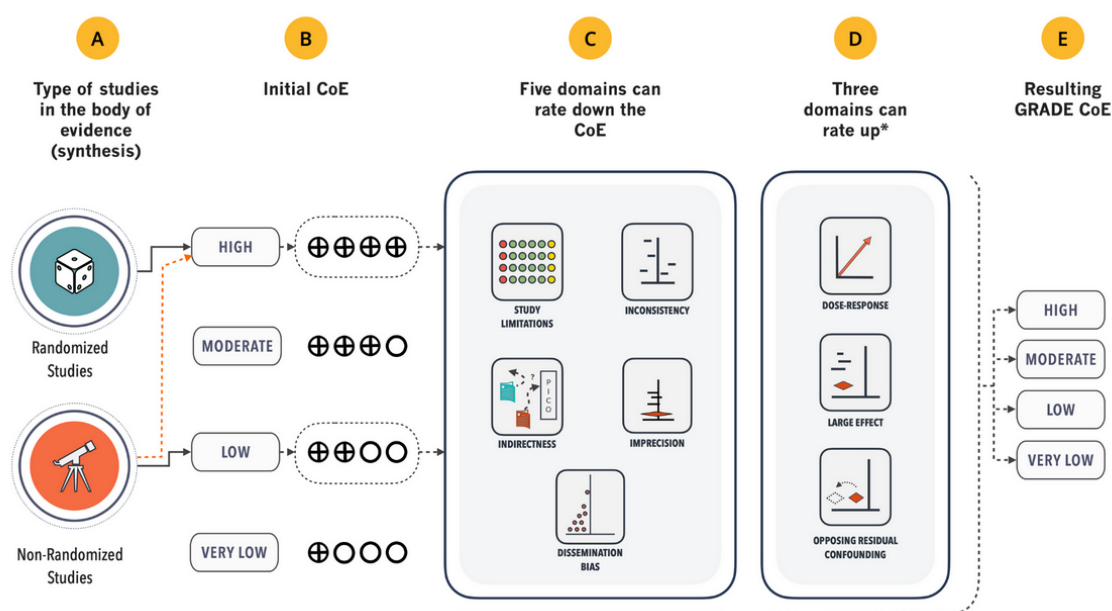
Certainty of the evidence is the "certainty that the true value of an estimate falls within a specific range, such as a small effect, or above or below a decision-making threshold, such as an important difference".

In the context of making recommendations, the certainty ratings reflect the extent of our confidence that the estimates of an effect are adequate to support a particular decision or recommendation. In the context of a systematic review, the ratings of the certainty of evidence reflect the extent of our confidence that the estimates of the effect are correct.

5.5.1.1 Four levels of evidence

GRADE categorises the certainty of the evidence in four levels: high, moderate, low, and very low. that are applied to a body of evidence, and not to individual studies (see resulting GRADE column E Figure 4). The definitions are specified in Table 9.

Figure 4 – The certainty of evidence by the GRADE approach



Source: see [website](#) ³²; Red dotted line represents non-randomised diagnostic accuracy studies which start at initial high certainty (see section 6.7).

Randomised trials start as high certainty evidence (column A), non-randomised studies (observational studies) as low certainty. The certainty of the evidence defined in GRADE means more than risk of bias (see column C study limitations). It also includes, and may be compromised by imprecision, inconsistency, indirectness of study results, and publication bias (or dissemination bias). In addition, several factors can increase our confidence in an estimate of effect. This general approach is summarised in Figure 4.

Table 9 – What it means: four levels of certainty of evidence

High certainty

There is high confidence that the true value of the estimate of interest is on one side of a threshold of interest or within a specific range.

Moderate certainty

There is moderately confidence that that the true value of the estimate of interest is on one side of a threshold of interest or within a certain range. The true value of the estimate may deviate slightly from the target of the certainty rating (i.e. may possibly fall in a different range).

Low certainty

There is low confidence that that the true value of the estimate of interest is on one side of a threshold of interest or within a certain range. The true value of the estimate may deviate from the target of the certainty rating (i.e. likely fall in a different range).

**Very low certainty**

There is very-low confidence that the true value of the estimate of interest is on one side of a threshold of interest or within a certain range. The true value of the estimate may deviate significantly from target of the certainty rating (i.e. probably fall in a different range).

5.5.1.2 Overall certainty of evidence

Each recommendation is defined by an overall certainty of evidence. When determining the overall certainty of evidence across outcomes:

- Consider only those outcomes that are deemed critical;
- If the certainty of evidence differs across critical outcomes and outcomes point in different directions — towards benefit and towards harm — the lowest certainty of evidence for any of the critical outcomes determines the overall certainty of evidence;
- If all outcomes point in the same direction — towards either benefit or harm — the highest certainty of evidence for a critical outcome, that by itself would suffice to recommend an intervention, determines the overall certainty of evidence. However, if the balance of the benefits and harms is uncertain, the grade of the critical outcome with the lowest certainty grading should be assigned.

5.5.1.3 GRADE and meta-analysis

GRADE relies on the judgement about our confidence in a (beneficial or adverse) effect of an intervention. Therefore, meta-analysis should be performed whenever possible.

It is possible that no pooled effect can or should be calculated if there is evidence of heterogeneity, be it clinical, methodological or merely statistical. GRADE has published some guidance on how GRADE can be applied in that case.³⁶

5.5.2 Study limitations, Risk of bias (Update 2025)**5.5.2.1 Risk of bias assessment of individual studies should be done using a validated assessment tool**

For the quality assessment of individual studies, specific tools were selected by the KCE (see section 4).

GRADE suggests that, based on the critical appraisal of individual studies, reviewers assess the overall risk of bias in individual studies as low or high. There is no definition on the number of high risk of bias items that merit rating a study as overall high risk of bias. This might be done for only one high risk category or item or require two or even more high risk categories or items to classify a study as high risk of bias. Thus, review teams may—and indeed do—mark overall bias differently.³⁷

Different outcomes from the same study may be at different risk of bias. For instance, risk of bias due to missing outcome data may be higher for an outcome that is difficult to follow for a long period (e.g., quality of life) than for an outcome that is easy to follow (e.g. survival). The same is true for risk of bias due to not blinding of outcome assessors: the impact of not blinding on risk of bias may differ across subjective (higher risk) versus objective (lower risk) outcomes.³⁷

Hence, it is advisable to report risk of bias assessment separately for different outcomes, or report transparently that the risk of bias is considered similar for all outcomes.

5.5.2.2 *Moving from individual risk of bias to a judgement about risk of bias across a body of evidence*

Moving from risk of bias criteria for each individual study to a judgement about rating down for risk of bias across a group of studies addressing a particular outcome presents challenges. GRADE suggests the following principles (summary in Table 10):

- First, in deciding on the overall certainty of evidence, one does not average across studies (for instance if some studies have no serious limitations, some serious limitations, and some very serious limitations, one does not automatically rate down by one level because of an average rating of serious limitations). Rather, judicious consideration of the contribution of each study, with a general guide to focus on the high-quality studies, is warranted.
- Second, this judicious consideration requires evaluating the extent to which each trial contributes to the estimate of magnitude of effect. This contribution will usually reflect study sample size and number of outcome events: larger trials with many events will contribute more, much larger trials with many more events will contribute much more. If studies with a high risk of bias dominate the body of evidence, reviewers should pool the results of all studies and rate down the certainty of evidence for risk of bias. However, when appreciable evidence from low risk of bias studies exists, reviewers should consider whether low and high risk of bias studies suggest similar or substantially different magnitudes of effect. If low and high risk of bias studies suggest substantially different interventions effects, reviewers should base inferences on only the low risk of bias studies as their best estimate of effect and not rate the certainty down for risk of bias. If results are not importantly different in low and high risk of bias studies, all studies can be included in the best estimation of intervention effects and one should not rate down the certainty of the evidence.³⁷ In addition, the expected direction of any bias influencing the results should be taken into consideration. If consideration of risk of bias would reinforce the conclusions, consequently there would be no reason to rate down for risk of bias.³⁷
- Third, one should be conservative in the judgement of rating down. That is, one should be confident that there is substantial risk of bias across most of the body of available evidence before one rates down for risk of bias.
- Fourth, the risk of bias should be considered in the context of other limitations. If, for instance, reviewers find themselves in a close-call situation with respect to two quality issues (risk of bias and, e.g. precision), we suggest rating down for at least one of the two.
- Fifth, notwithstanding the first four principles, reviewers will face close-call situations. They should both acknowledge that they are in such a situation, make it explicit why they think this is the case, and make the reasons for their ultimate judgement apparent.



Table 10 – GRADE’s approach for grading down the certainty of the evidence for risk of bias in a body of evidence

Extent of risk of bias	Risk of bias within a study	Risk of bias across studies	Interpretation across studies
No serious limitations, do not downgrade	Low risk of bias for all key criteria	Most information is from studies at low risk of bias	High certainty: the true effect lies close to that of the estimate of the effect
Serious limitations, rate down one level (i.e. from high to moderate certainty)	Crucial limitation for one criterion or some limitations for multiple criteria sufficient to lower ones confidence in the estimate of the effect.	Most information is from studies at moderate risk of bias.	Certainty of evidence reduced from high- to moderate certainty evidence; the true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different.
Very serious limitations, rate down two levels (i.e. from high to low certainty or moderate to very low)	Crucial limitation for one or more criteria sufficient for substantially lowering ones confidence in the estimate of the effect.	Most information is from studies at high risk of bias.	Certainty of evidence reduced from high – to low certainty evidence: the true effect may be substantially different from the estimate of the effect.

Source Guyatt ³⁸

5.5.3 Inconsistency (Update 2025)

When performing a meta-analysis, exploring and judging inconsistency (also named heterogeneity) is probably the most difficult step. A number of rules are presented, but a full explanation can be found in the [Cochrane Handbook](#) (chapters 10.10 and 10.11; Version 6.5, 2024). A KCE expert should have a good knowledge and the ability to apply the Cochrane guidance, if this is not the case, the help from a more expert person should be asked.

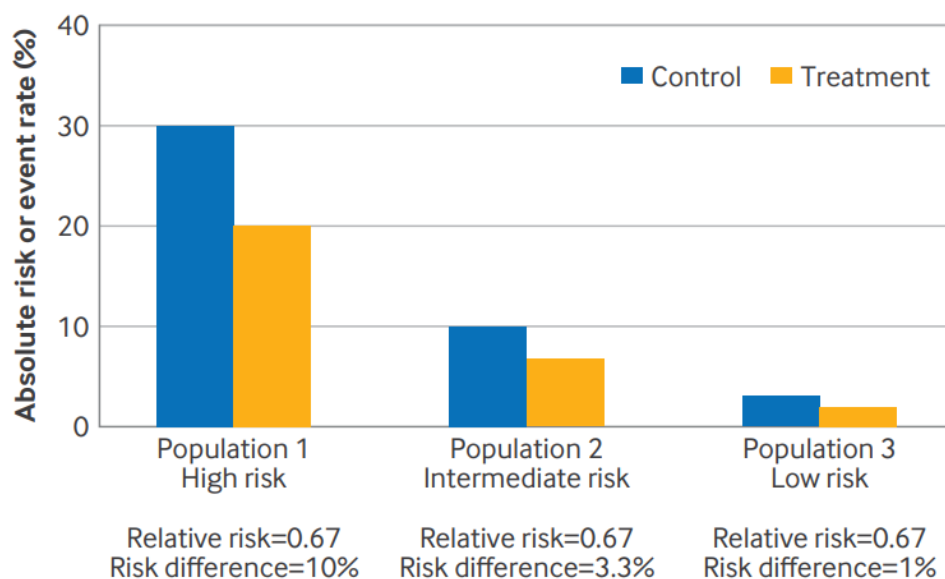
In GRADE, inconsistency refers to unexplained variability in results across studies. More particularly, GRADE focuses on inconsistency that is sufficiently large that, depending on which of the varying results represents the truth, inferences for clinical practice would differ. Importantly, the GRADE inconsistency domain does not address variability in study design (PICO), but rather variability in study results.³⁹

In general terms, when inconsistency in effects is observed, there may be several reasons for this inconsistency. These include random error and differences in population, interventions, comparator or outcome.

Hypotheses about the reasons for the observed inconsistency may or may not explain the differences in effects. When variability in results exists, but investigators fail to identify a plausible explanation arising from differences in patient characteristics, interventions, comparators or outcomes (e.g. length of follow-up), the certainty of the evidence should be graded down by one or two levels, depending on the magnitude of the inconsistency in the results.⁴⁰

Note that relative effects very seldom show important differences across groups. However, despite relative effects being similar in groups with different baseline risks, absolute effects may differ substantially (Figure 5). In case of similar relative effects but different baseline risks, risk differences (absolute effects) for each relevant patient group should be reported in the summary of findings tables (by applying the relative risks to the respective baseline risks) and subsequent recommendations may differ for low, medium or high-risk patients.⁴⁰

Figure 5 – Constant relative risk for varying baseline risk resulting in differing absolute risks



Source Core GRADE 3⁴⁰

GRADE suggests rating down the quality of evidence only if large inconsistency (heterogeneity) in study results remains after exploration of *a priori* hypotheses that might explain heterogeneity (see 5.5.3.4).

5.5.3.1 Inconsistency versus heterogeneity

GRADE uses inconsistency and heterogeneity rather interchangeably; nevertheless, the following should be considered.

- A heterogeneity in effect – where it can be assumed / explained that it is randomly distributed – may be due to random variation in the effect amongst studies. To properly address this, the pooled effect should be calculated using random modelling (RevMan uses the DerSimonian and Laird random effects model, but other techniques, such as Bayesian and maximum likelihood, are often used as well). An important condition for the use of these techniques is that it must be plausible that the heterogeneous effect is randomly distributed, which is not always easy to verify. DerSimonian, Laird and maximum likelihood methods have an additional assumption that the effect is normally distributed, while with Bayesian techniques another distribution can be used as well. The studies in this case cannot be considered as inconsistent, and the heterogeneity is accounted for by the larger confidence interval, so no downgrading is needed. Note that if the heterogeneity statistic Q is less than or equal to its degrees of freedom (so if $I^2 = 0$), DerSimonian gives results that are numerically identical to the (non-random effects) inverse variance method.
- If heterogeneity is important for one reason or another, but all estimates point in the same direction, e.g. a strong or very strong effect of the intervention, then one should not necessarily grade down for inconsistency but make a judgement on the plausibility of the study results.

5.5.3.2 Judging inconsistency (Update 2025)

In a first step, for binary outcomes, inconsistency should be judged based on relative effects, such as risk ratios or hazard ratios. For dealing with continuous outcomes, inconsistency judgements should focus on absolute effects such as mean differences.⁴⁰

Next, a small number (three or fewer) of *a priori* hypotheses should be specified that might explain (large) inconsistencies. This is done during the preparation phase of the analysis (protocol stage). The

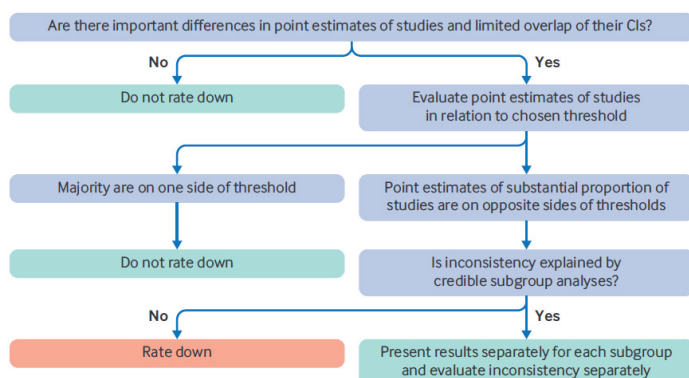


choice of hypotheses should be based on previous evidence or a thorough understanding of the underlying biology and should include the direction of the hypothetical subgroup effect.

GRADE identifies three criteria from forest plots for assessing serious inconsistency in results, and reviewers should consider rating down for inconsistency when:

- Point estimates vary widely across studies;
- Confidence intervals show minimal or no overlap;
- Point estimates don't lie largely on the same side of a chosen threshold.

Figure 6 – Judging inconsistency with Core GRADE



Source Core GRADE 3 ⁴⁰

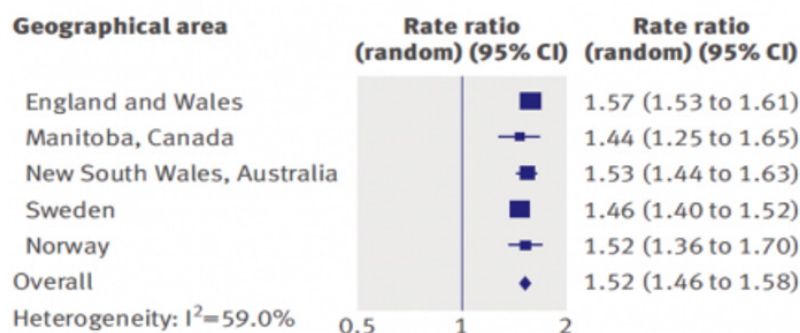
In addition, one statistical assessment criterion is suggested: the I^2 describing the percentage of the variability in effect estimates due to heterogeneity rather than sampling error (chance).

In the past, rigid criteria were used to judge heterogeneity, e.g. an I^2 of >50% was commonly applied. This improves the consistency in judgements, but one risks to be consistently wrong. All statistical approaches have their limitations, and the result should be seen in the context of a subjective examination of the variability in point estimates and the overlap in CIs.

What is a large I^2 ? One set of criteria would say that an I^2 of less than 40% is low, 30 to 60% may be moderate, 50 to 90% may be substantial, and 75 to 100% is considerable. Note the overlapping ranges and the equivocation ("may be"): an implicit acknowledgment that criteria are both arbitrary and uncertain. When individual study sample sizes are small, point estimates may vary substantially, but because variation may be explained by chance, I^2 may be low. Conversely, when study samples sizes are large, a relatively small difference in point estimates can yield a large I^2 .

Example of how I^2 can be misleading

Consider this meta-analysis, both the graph and I^2



All point estimates are around 1.5 and the estimates form a near straight line. Nevertheless, I^2 suggests substantial heterogeneity (P-value Q-test=0.04). Conclusion: I^2 is high because the underlying study sample sizes are large. Reliance on the strict quantified criteria for I^2 would lead to a wrong or even an absurd conclusion to rate down.

5.5.3.3 Other considerations (Update 2025)

- Risk differences (i.e. absolute risk reductions) in subpopulations tend to vary widely. Relative risk (RR) reductions, on the other hand, tend to be similar across subgroups, even if subgroups have substantial differences in baseline risk. GRADE considers the issue of difference in absolute effect in subgroups of patients, much more common than differences in relative effect, as a separate issue. When easily identifiable patient characteristics confidently permit classifying patients into subpopulations at appreciably different risk, absolute differences in outcome between intervention and control groups will differ substantially between these subpopulations. This may well warrant differences in recommendations across subpopulations. However, differences in absolute effects only without differences in the relative effect does not warrant rating down for inconsistency.
- Rate down for inconsistency, not up for consistency.
- Even when there is heterogeneity in effect, one must evaluate if the heterogeneity affects your judgement on clinical effectiveness, e.g. when there are large differences in the effect size, but when the estimations point to the same direction (all beneficial or all harmful)(see third criterion from forest plots).
- Rating down twice for inconsistency should not be considered.⁴⁰

5.5.3.4 Inconsistency and subgroup analyses (Update 2025)

Beware of subgroup analyses. The warning below originates from the Cochrane Handbook (chapter 10.11.2, Version 6.5 2024 see [website](#)). When confronted with this, consult at least a second opinion of a knowledgeable person.

Subgroup analyses involve splitting all the participant data into subgroups, often so as to make comparisons between them. Subgroup analyses may be done for subsets of participants (such as males and females), or for subsets of studies (such as different geographical locations). Subgroup analyses may be done as a means of investigating heterogeneous results, or to answer specific questions about particular patient groups, types of intervention or types of study. Findings from multiple subgroup analyses may be misleading. Subgroup analyses are observational by nature and are not based on randomised comparisons (an exception is when randomisation is stratified within these subgroups). False negative and false positive significance tests increase in likelihood rapidly as more subgroup analyses are performed (this is due to the multiple testing problem: if you perform a



significant test frequently enough, you are likely to find by chance a statistically significant result). If findings are presented as definitive conclusions, there is clearly a risk of patients being denied an effective intervention or treated with an ineffective (or even harmful) intervention. Subgroup analyses can also generate misleading recommendations about directions for future research that, if followed, would waste scarce resources.

If a subgroup effect is suspected, GRADE suggests to assess the credibility of the subgroup effect by using a validated checklist, for example the Instrument for assessing the Credibility of Effect Modification Analyses (ICEMAN).⁴¹ An example is written out in GRADE guidance 36.³⁹ Before embarking on a full assessment, a test of interaction should be performed; If the P-value for the test is 0.1 or higher, no further exploration of subgroup effect is warranted.³⁹

When, according to the assessment, the credibility in a subgroup effect is low, the overall estimate should be used. If credibility in a subgroup effect is moderate to high, separate effects should be used in the certainty of evidence rating, the summary of findings tables and the development of recommendations.³⁹

The GRADE approach to addressing subgroup analyses is summarised in Table 11.

Table 11 – GRADE's approach to addressing subgroup analyses

1 For the PICO	Consider possible absolute and relative subgroup effects that may be present
2 For absolute effects	a) Seek observational studies that define patient groups at varying baseline risk b) For populations in whom easily identifiable patient characteristics identify substantial differences in baseline risk, present in summary of findings tables absolute benefits for each subpopulation c) For guidelines, consider whether populations at varying baseline risk warrant different recommendations
3 For relative effects	a) Generate a priori subgroup hypotheses, including postulated direction of effects, ideally based on prior evidence. Limit the number of subgroups to investigate, ideally no more than three b) When evidence is available, check for possible subgroup effects, ideally focusing on within-study comparisons c) The check for possible subgroup effects, begin with calculation of a test of interaction. If the P-value for the test is 0.1 or higher, no further exploration of subgroup effect is warranted. In such case, observed inconsistency cannot be sufficiently explained by subgroup effects and rating down should be considered. d) If the P-value of test of interaction is smaller than 0.1, apply ICEMAN or, alternatively one of the available checklists for determining the credibility of the subgroup analysis e) If the credibility is low or very low, present in summary of findings tables the overall estimate. If credibility is moderate or high, each subgroup should have its own summary of findings table with relative and absolute effects and certainty rating. f) For guidelines, consider whether different recommendations are warranted for each subgroup as it is very likely that different recommendations will be needed.

Source: GRADE guidance 36, Guyatt et al. 2023³⁹ ; ICEMAN instrument⁴¹

5.5.3.5 Quantitative assessment of inconsistency using decision thresholds (optional) (Update 2025)

GRADE developed two measures that can complement the above described approaches to assess inconsistency: the Decision Inconsistency index (DI) and the Across-Studies Inconsistency index (ASI).⁴² Definite guidance on cut-off points and on how to use the indices in rating down for inconsistency however is still under development.

The Decision Inconsistency index (DI)

The DI quantifies overall inconsistency from a decision point of view.

Consider a comparison for which an outcome-specific decision threshold (DT) has been established so that:

- If the effect size (ES) is larger than the DT, the intervention would be associated with at least small harms.
- If the $ES < -DT$, the intervention would be associated with at least small benefits.
- If $-DT \leq ES \leq DT$, the intervention would be associated with a trivial or no effect.

In the example, there are 3 'decision categories', but the DI can be calculated for as many interpretation categories as desired (e.g. including small, moderate and large effects). The ES should be presented as an absolute effect. The DI ranges between 0 and 1 with 0 indicating no DT-related inconsistency and 1 indicating maximum DT-related inconsistency. In case of no inconsistency, all effect sizes will fall in the same category. In case of 'complete inconsistency', there will be a perfectly even distribution of the effect sizes across the three interpretation categories.

The Across-Studies Inconsistency index (ASI)

The ASI measures across-study inconsistency considering decision interpretation categories. To assess across-studies inconsistency, the observed number of samples per interpretation category for each study is compared with the number of samples (per interpretation category for each study) if the proportion of samples per interpretation category had been the same for all studies and equal to the overall proportion. The ASI values can range between 0 and 1, with 0 indicating no across-studies inconsistency and 1 indicating complete across-studies inconsistency.

The formulas for calculating the DI and the ASI are available in the publication by Sousa-Pinto et al.⁴² The indices can also be calculated using the app available at the [website](#).

5.5.4 Indirectness (Update 2025)

Two types of indirectness can be distinguished: indirectness due to indirect comparisons and indirectness due to PICO issues.

Direct evidence comes from research that directly compares the interventions in which we are interested when applied to the populations in which we are interested and measures outcomes important to patients. Indirect evidence is useful when clinicians must choose between interventions that have not been tested in head-to-head comparisons. Making comparisons between treatments under these circumstances requires specific statistical methods. This is further presented in section 5.6. Grade continues to develop specific guidance for rating the certainty of evidence from network meta-analysis (see section 5.6).⁴³⁻⁴⁵



5.5.4.1 Indirectness due to PICO issues (Update 2025)

Evidence can be indirect due to PICO elements in one of four ways:

- First, patients may differ from those of interest (the term applicability is often used for this form of indirectness). Examples that may lead to downgrading are guidelines for treatment in children when only evidence in adults is available or changes over time in virus variants during the covid pandemic. Indirect evidence for harms (side effects) from other populations often does not require downgrading, as it often can be assumed that the adverse effects are similar irrespective of the illness for which it is used.⁴⁶
- Secondly, the intervention tested may differ from the intervention of interest. Decisions regarding indirectness of interventions depend on an understanding of whether biological or social factors are sufficiently different that one might expect substantial differences in the magnitude of effect. Typical examples are differences in dose or duration of the intervention or route of administration. Other possible reasons to grade down for indirectness related to the intervention are:
 - Differences in skill level of providers of the intervention, such as in educational, surgical, physiotherapy and psychological interventions;
 - Inadequate description of the intervention, for example in cardiac rehabilitation (reduced certainty that the estimated effect of the intervention will be achieved when implemented due to insufficient information about the intervention);
 - Differences in adherence: as an individual patient, you may be interested in the effect of a screening intervention when fully adhering to the screening schedule. On the other hand, policy makers may be interested in the effect of a screening programme with realistic participation rate. Studies of a behavioural intervention that is difficult to follow for many patients in real life, may provide only indirect evidence if they include exclusively highly motivated and trained patients.
 - Change of intervention technology: trials including older versions of devices or technology may only provide indirect evidence for the use of newer versions. When reviewing the evidence, two strategies can be followed: excluding studies of obsolete systems or including studies of obsolete systems and lowering the certainty of evidence for indirectness.
 - It is also suggested that switching treatments in clinical trials (cross-over) can be reason for rating down for indirectness. Authors should judge the appropriateness of the cross-over when considering to rate down for indirectness. Furthermore, if the effect (of superiority) would only be greater without the switch in treatment, one should not rate down for indirectness.⁴⁶
- Thirdly, the comparator may differ from the target PICO. Examples include the use of placebo or inferior older alternatives when a clinically relevant active comparator is available and variations in standard of care in different regions.⁴⁶
- Fourthly, outcomes may differ from those of primary interest, for instance, surrogate outcomes that are not themselves important, but measured in the presumption that changes in the surrogate reflect changes in an outcome important to patients.⁴⁶

5.5.4.2 Indirectness versus inconsistency (Update 2025)

When developing clinical guidelines, often a systematic review is performed to identify evidence answering a 'broad' PICO' question without strict eligibility criteria. When the PICO is developed, also a priori hypotheses for possible explanations of inconsistency should be generated (see 5.3.2.).

If subsequently, evidence is found for all subgroups of interest, differences in effect between subgroups can be tested and evaluated for credibility as described earlier for assessing whether to rate down for inconsistency. However, when only evidence is identified for one subgroup that is

different from the original PICO, differences in effect between subgroups cannot be assessed and rating down for indirectness should be considered.⁴⁶

5.5.5 Imprecision (Update 2025)

Before assessing whether or not to rate for down for imprecision, it is important to decide what the target of the certainty rating is. The focus of Core GRADE is on two questions:⁴⁷

- Whether there is an effect compared with an alternative. In that case, the target is the null threshold (“How certain are we that the effect is different from no effect?”).
- Whether there is an effect that is large enough for patients. In that case, the target is the minimal important difference (MID) (How certain are we that there is an effect that is equal to or larger than the MID?). (see 5.5.5.1)

It is also possible to define thresholds for small, moderate and large effects, and considering one of these thresholds as the target for the certainty rating. The thresholds should be a result of combining the relative importance of the outcome and an absolute effect estimate (see 5.5.5.1). Identifying and defining thresholds for health outcomes is part of the first step when judging imprecision.^{7, 47}

GRADE suggests that examination of 95% confidence intervals (CIs) provides the optimal primary approach to decisions regarding imprecision. Rating should be based on absolute effects and may differ for groups with a different baseline risk. For binary outcomes, the best estimate of the risk difference and its 95%CI is used. Absolute effect estimates will have narrower confidence intervals in a population with a very low baseline risk than in a population with a high baseline risk, even if the relative risk and its confidence interval are the same.⁷

As a general principle GRADE recommends to consider the rating down for imprecision if a recommendation or clinical course of action would differ if the upper versus the lower boundary of the CI represented the truth. In other words, rating down certainty for imprecision should be considered when the CI crosses the chosen threshold. If the certainty is rated that the effect is beyond or below one particular threshold, one needs to use that threshold to rate imprecision. If the MID is chosen as the threshold and the point estimate is greater than the MID, reviewers will rate the confidence that there is an important effect. If the point estimate is less than the MID, reviewers will rate the certainty that the true effect is unimportant. Thus, where the point estimate falls in relation to the chosen threshold determines the target of the certainty rating.^{7, 33, 47}

One can consider to rate down by two or three levels when the CI for the absolute effect crosses respectively two or three thresholds, for example if the CI includes both important benefit and harm.⁷ CORE grade also refers to the plain language statements (which one suits best) when deciding whether to rate down one or twice for imprecision.⁴⁷

As the judgement of imprecision is based on examination of the 95%CI, it is important that the meta-analysis is correctly performed. An important choice that has to be made is whether a fixed effect or random effect statistical model is used. Some easily accessible information can be found in the Core GRADE paper (appendix 2).^{47, 48}

5.5.5.1 Clinical decision threshold, minimally important difference and categories of magnitude of effect (Update 2025)

As we explained before, it is good to discuss the clinical decision thresholds and try to determine them on beforehand.

The MID or the small effect threshold represents the smallest difference in an individual outcome that people judge as important (either in general or the target population). Additional ranges of no or trivial, moderate and large effects facilitate decision making using the evidence-to-decision frameworks.³⁴

GRADE further explains how the thresholds can be defined for dichotomous outcomes and for continuous outcomes.



DICHOTOMOUS OUTCOMES

Thresholds for desirable and undesirable health effects will usually be symmetric and depend on the utility or the value assigned to the outcomes. In general, thresholds for critical outcomes are expected to be smaller than those for important outcomes. Empirical thresholds for small, moderate and large effects are available in the literature (Morgano et al. 2025 JCE 2025;179:111639) and can be used together with utilities to calculate thresholds.^{7, 49} An alternative approach to calculations is using thresholds based on informed guesses by the stakeholders or based on thresholds set in prior HTAs or guidelines. It is important to always be transparent on how the threshold was set.⁷

If it is not possible to determine a threshold, then GRADE suggests to use a RRR or RRI greater than 25%. This 25% threshold is a good and pragmatic starting point, but other factors should be taken into account as well, such as the absolute risk reduction.

CONTINUOUS OUTCOMES

If empirical estimates for small, moderate and large effects for continuous outcomes are available in the literature, they should be used to define relevant thresholds.

If such empirical estimates are not available, it is suggested to standardize the effect size, express it as a standardized mean difference and use an interpretation of the effect size offered by Cohen, also known as 'Cohen's d'. According to this interpretation, an effect size or SMD of around:^{7, 50}

- 0.2 is considered a small effect;
- 0.5 is considered a moderate effect;
- 0.8 or higher is considered a large effect.

There are also other methods to standardise effect sizes, such as Glass Δ or hedges d, depending on the choice of the estimator of the standard deviation.

5.5.5.2 *Relative versus absolute effects*

The recommended output of a meta-analysis is a relative risk or odds ratio as they are more stable across different baseline risks, but absolute reductions are more useful for clinical decisions. The valid way of estimating the absolute risk reduction is to multiply the relative risk from the meta-analysis with the most plausible estimate or range of estimates of the baseline risk, this is the risk in the assumed control group, often called the assumed control risk.

Assumed control risk (ACR)

As relative risks are more stable, absolute risk reductions are best calculated applying a relative risk reduction to an assumed control risk. Determining the assumed control risk is not always easy and straightforward. The best estimates usually come from Belgian data, e.g. for cancer the data from the cancer registry are a valuable tool. When this information is not available one can use expert opinion or take the baseline risk from studies in the literature. When different studies exist with varying baseline risk sometimes the median control risk is taken (this is what GRADE pro does by default), but it may be better to report a range of plausible values.

Appendix 3 of the Core GRADE 2 paper on imprecision provides some examples.⁴⁷

GRADE guidelines 27 provide guidance on how to calculate absolute effects for time-to-event outcomes.⁵¹

5.5.5.3 Dealing with fragility: optimal information size (Update 2025)

When the CI crosses the threshold or thresholds of interest, certainty of the evidence will be rated down for that reason, and there is no need to consider the sample size. If the CI does not cross the threshold and the effect is large, the clinical decision threshold criterion is insufficient (CI is fragile, not reliable) to address issues of precision. The reason is that small numbers of events may render the results fragile.⁴⁷

Example

Consider a randomised trial of β blockers in 112 patients undergoing surgery for peripheral vascular diseases that fulfilled pre-planned O'Brien-Fleming criteria for early stopping. Of 59 patients given bisoprolol, 2 suffered a death or nonfatal myocardial infarction, as did 18 of 53 control patients. Despite a total of only 20 events, the 95%CI around the RR (0.02 to 0.41) suggests a large treatment effect, even for the upper limit of the CI. The CI suggests that the smallest plausible effect is a 59% RR reduction. Were this the case, we would immediately administer treatment.

However, there are reasons to doubt the estimate of the magnitude of effect of this trial. First, it is much larger than what we might expect on the basis of β blockers' effects in a wide variety of other situations. Second, the study was terminated early on the basis of the large effect. Third, concluding that an RRR less than 59% is implausible on the basis of only 20 events violates common sense: intuitively, we have a sense of the fragility of these results. Our intuitive scepticism is justified: if one moves just five events from the control group to the intervention group, the results lose their statistical significance, and the new point estimate (a RRR of 52%) is outside the original CI.

The reasoning above suggests the need for, in addition to CIs, another criterion for adequate precision. GRADE suggests the following: if the total number of patients included in a systematic review is less than the number of patients generated by a conventional sample size calculation for a single adequately powered trial, consider the rating down for imprecision. Authors have referred to this threshold as the "optimal information size" (OIS). Many online calculators for sample size calculation are available, you can find one simple one at the [website](#).

When performing such a sample size calculation, Core GRADE suggests to use typical error rates for α and β (0.05 and 0.20) and a modest relative risk reduction, for example 20% or 25% for binary outcomes. For continuous outcomes, one can use the MID and the standard deviation from one of the existing studies. As an alternative, as a rule of thumb, Core GRADE suggests not to have concerns about imprecision (thus not rating down) if each group includes more than 400 patients (total sample size 800).⁴⁷

The update full GRADE guidance refers to the review information size (RIS). When the observed absolute effect is large, one might want to determine if the effect is implausibly large, e.g. based on randomly high effect sizes. To do so the RIS can be calculated by calculating the sample size for hypothetical trials to test the difference between the observed effect and the thresholds for large, moderate and small effects. Note that the calculated RIS will depend on the control event rate. GRADE's guidance suggests that if the actual RIS is smaller than that calculated for the large, moderate or small effect thresholds, the corresponding rating of imprecision should be lowered by three, two or one level(s) respectively. A worked example can be found in GRADE guidance 35.⁷

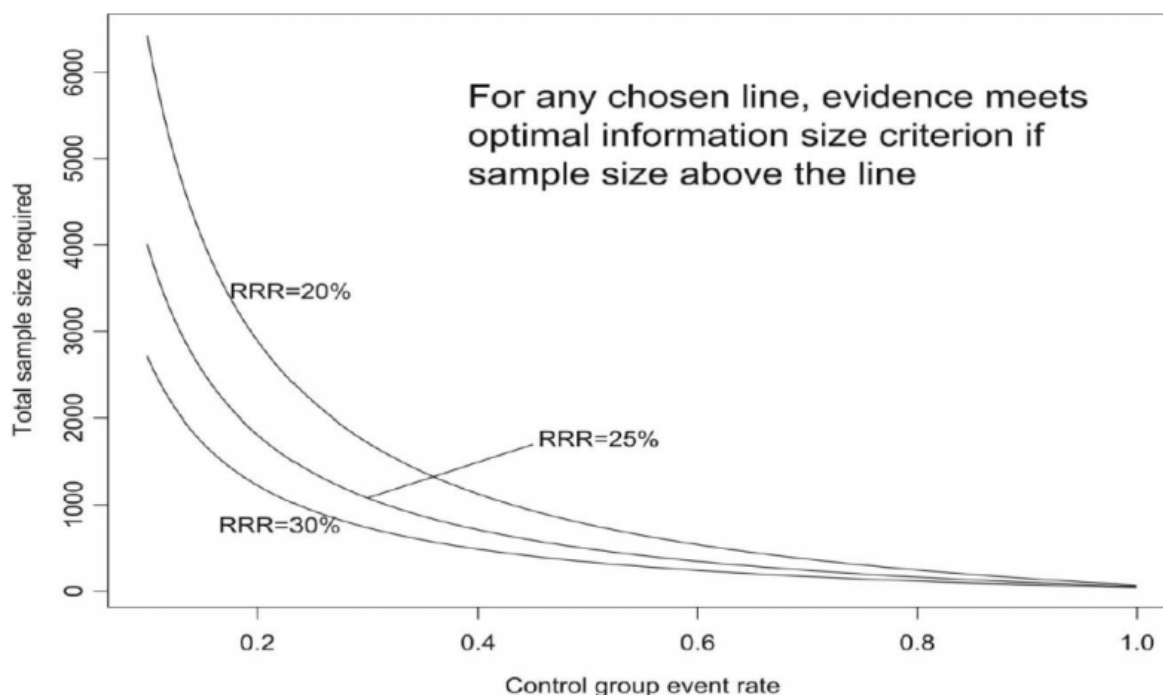
Furthermore, review and guideline authors can also consult a figure to determine the OIS. Figure 7 presents the required sample size (assuming α of 0.05, and β of 0.2) for RRR of 20%, 25%, and 30% across varying control event rates. For example, if the best estimate of control event rate was 0.2 and one specifies an RRR of 25%, the OIS is approximately 2 000 patients.

The choice of RRR is a matter of judgement. The GRADE handbook suggested using RRRs of 20% to 30% for calculating the OIS, but there may be instances in which compelling prior information would suggest choosing a larger value for the RRR for the OIS calculation.



Beware, however, not to base your sample size on the RRR of minimally clinical importance, a practice that is suitable for sample size calculations when you set up studies, but not for judging fragility, because it leads to paradoxes: if expected effects are considerably larger than what is clinically important because the clinically important effect is small (e.g. a small effect on mortality in children would be considered important), you would risk to grade down without good reasons because the required sample size would be too large. Note that the OIS helps judging the stability of the CIs, and not whether the study was large enough to detect a difference.

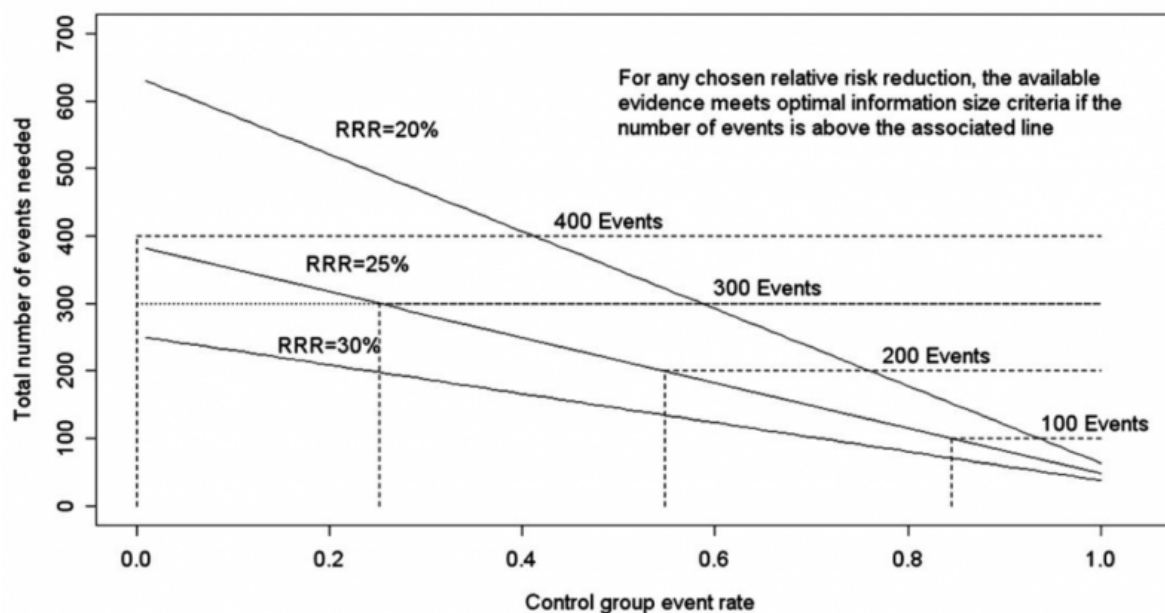
Figure 7 – Optimal information size by control group event rate



Source: Guyatt et al. 2025 Core GRADE 2 (Appendix 5)⁴⁷

Power is, however, more closely related to number of events than to sample size. Figure 8 presents the same relationships using total number of events across all studies in both treatment and control groups instead of total number of patients. Using the same choices as in the prior paragraph (control event rate 0.2 and RRR 25%), one requires approximately 325 events to meet OIS criteria.

Figure 8 – Total number of events needed by control group event rate



Source: Guyatt et al. 2025 Core GRADE 2 Appendix 5⁴⁷

Calculating the OIS for continuous variables requires specifying:

- probability of detecting a false effect – type I error (α ; usually 0.05)
- probability of detecting a true effect – power (usually 80% [power = 1 – type II error; β ; usually 0.20])
- realistic difference in means (Δ)
- appropriate standard deviation (SD) from one of the relevant studies (we suggest the median of the available trials or the rate from a dominating trial, if it exists).

For continuous variables we should grade down when the total population size is less than 400 patients per group (or a total sample size of 800; a threshold rule-of-thumb value; using the usual α and β , and an effect size of 0.2 SD, representing a small effect). In general, a total number of events of more than 400 guarantees the stability of a confidence interval.⁴⁷

Finally, one can rely on the ratio of the upper to the lower boundary of the CI for an Odds ratio or Risk ratio. The upper and lower boundary of the CI may indicate whether the sample size in the meta-analysis is less than the OIS if their ratio is greater than 2.5 for odds ratio and 3 for risk ratio.⁴⁷

5.5.5.4 Low event rates with large sample size: an exception to the need for OIS

When event rates are very low, CIs around relative effects may be wide, but if sample sizes are sufficiently large, it is likely that prognostic balance has indeed been achieved, and rating down for imprecision becomes inappropriate. Unfrequently occurring adverse events are a typical example.

Example

Consider a systematic review of artemether-lumefantrine versus amodiaquine plus sulfadoxine-pyrimethamine for treating uncomplicated malaria. For serious adverse events (SAEs), the authors calculated a RR of 1.08 (95% CI: 0.56 to 2.08). They judged this CI sufficiently wide to rate down quality two levels (from high to low) for imprecision.

There were, however, only 34 SAEs in over 2 700 patients, corresponding to event rates of 1.2% and 1.3% in the two groups. Of these events, 2 were deaths, 2 severe anaemia and the remainder febrile seizures



and elevated liver function. In absolute terms, the difference between groups is 1 event per thousand patients with a CI from 6 in 1000 fewer to 14 in 1000 more. Particularly considering that very few of these SAEs were associated with long-term morbidity, focusing on the CI around absolute effects would lead to one to reject rating down quality two levels for imprecision, and possibly not rate down for imprecision at all.

The decision regarding the magnitude of effect that would be important is a matter of judgement. When control rates are sufficiently low, CIs around relative effects can appear very wide, but CIs around absolute effects will nevertheless be narrow. Thus, although one would intuitively rate down for imprecision considering only the CI around the relative effect, consideration of the CI around the absolute effect may lead to an appropriate conclusion that precision is adequate. Note that the inference of unimportance requires a low incidence of events over the desirable duration of follow-up; short follow-up will generate a low incidence of events that may be misleading.

5.5.5.5 Summary GRADE imprecision assessment (Update 2025)

In a simplified way, the steps to follow are summarised below

Step 1: Define your outcome as dichotomous or continuous.

Step 2: Set thresholds for the outcome under consideration, in absolute effects

Thresholds for a small, moderate and large effect are defined. If Core GRADE is used, only the MID (or small effect) is defined. Thresholds can be defined by calculations, medical literature or informed guesses by experts. Methods should be transparently reported.

Step 3: Choose the target of the certainty rating (i.e. decide if you rate if the effect lies beyond or below a threshold and which threshold).

Step 4: Calculate the absolute effect estimates for the body of evidence if not yet done so.

Step 5: Judge whether to rate down or not based on the 95%CI.

The number of levels to rate down depends on the number of thresholds that are crossed.

Core GRADE suggests to also take the most suitable evidence statement into account.

Step 6 (optional): if the effect is large and the certainty of the evidence is not rated down based on the 95%CI, take the RIS (or OIS) into account.

Calculate the required sample size for small, moderate or large effects for the outcome of interest. If the actual RIS is smaller than that calculated for the large, moderate or small effect thresholds, the corresponding rating of imprecision should be lowered by three, two or one level(s) respectively

Calculate the OIS when applying Core GRADE and rate down if the total number of events or participants is below the OIS. Alternatively, figures can be consulted or a rule of thumb of 800 participants in total (400 in each group) can be used.

5.5.6 Publication bias (Update 2025)

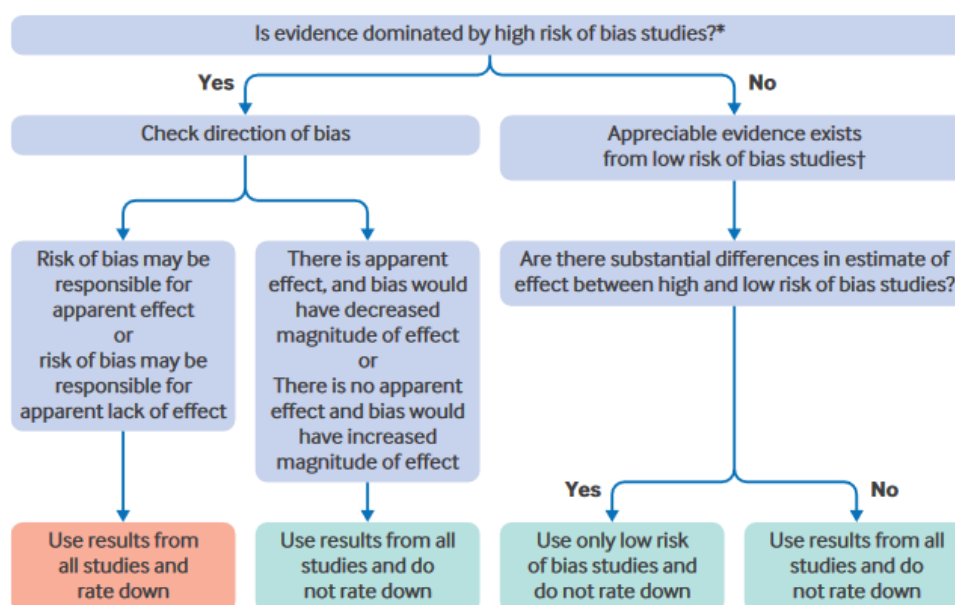
Publication bias is a systematic underestimate or an overestimate of the underlying beneficial or harmful effect due to the selective publication of studies based on their results. Publication bias arises when investigators fail to report studies they have undertaken (typically those that show no effect). Authors should consider rating down for publication bias when available evidence comes from a number of small studies, most of which have been commercially funded, because of the concern about the impact of industry sponsorship on selective publication.

Trials registries are in principle compulsory now and can play a major role in detecting selective reporting. Guidance on how to search trial registries can be found in the publication by [Hunter et al 2022 in the BMJ](#).⁵² A number of approaches based on the examination of the pattern of data are available to help assess publication bias. The most popular of these is the funnel plot. Funnel plots can be used to test publication bias by using visual inspection of asymmetry. If the funnel plot is asymmetrical with a missing quadrant of small studies with negative results, publication bias represents a plausible explanation. Statistical approaches, such as the Egger's regression test are also available but should only be used if the meta-analysis includes 10 or more studies.³⁷

All methods, however, have substantial limitations and authors of reviews and guideline panels must often guess about the likelihood of publication bias. Again, there is no substitute for judgement. Because of this uncertainty, it is suggested to use the terminology 'undetected' or 'strongly suspected' publication bias depending on the findings.³⁷

The steps to be followed when deciding whether or not to rate down for publication bias are summarised in Figure 9.

Figure 9 – Flow chart depicting process of deciding whether to rate down certainty of evidence for publication bias



Source: Guyatt 2024, Core GRADE 4.³⁷

Note that selective reporting of outcomes should be dealt with in the quality assessment of the individual studies (see section 4.3, 4.4, and 4.5).



5.6 GRADE for network meta-analysis (Update 2025)

5.6.1 Introduction

Rating the certainty in estimates from a network meta-analysis is a complex process asking for a good understanding of the general GRADE principles. In case of doubt or when experience with GRADE is limited, it is advised to consult a more experienced GRADE assessor. Table 12 provides a glossary with the main concepts covered in this chapter.

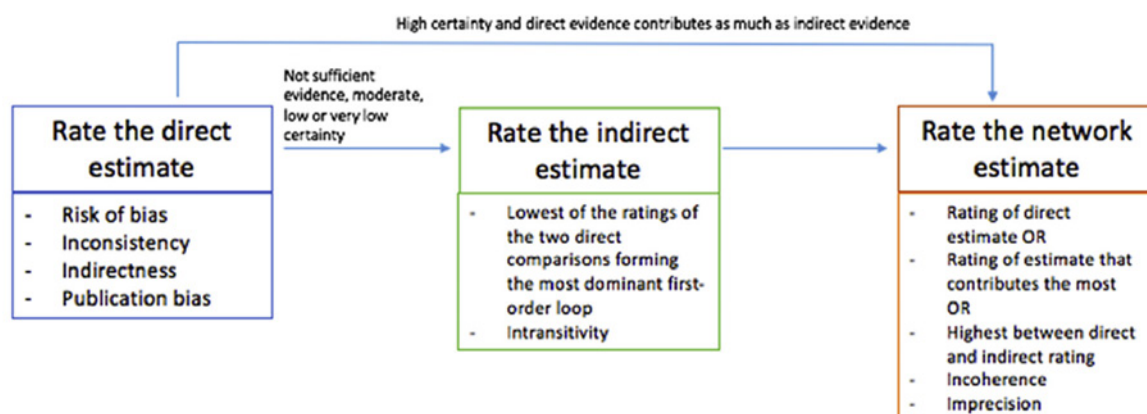
Where in a conventional meta-analysis the application of the GRADE approach concerns only one direct pairwise comparison (for example intervention A vs. intervention B), in a network meta-analysis (NMA) multiple direct and indirect comparisons can contribute to the final assessment of the certainty in estimates for a specific pairwise comparison. The least complex network would consist of two direct comparisons (between an intervention A and C and between an intervention B and C) and one indirect comparison (between an intervention A and B). This is an example of a first order loop, where only one additional intervention (i.e. C) is involved to build a comparison between two interventions (i.e. A and B). Second order loops would involve two additional interventions

In their guidance about rating the certainty in estimates from a network meta-analysis the GRADE Working Group pointed out three major steps that should be applied to each separate pairwise comparison of interest (see Table 12).^{44, 45} This flow is also depicted in Figure 10.⁴⁴

Table 12 – Process for rating the certainty of a network estimate

Steps	Perform the following
Step 1: <i>Rate the certainty of the direct estimate</i>	• Use the general GRADE principles • Focus on risk of bias, inconsistency, indirectness and publication bias; (not imprecision) • If assessed to be of high certainty, and if the direct evidence contributes to the network estimate as much as the indirect evidence, step 2 can be skipped
Step 2: <i>Rate the certainty of the indirect estimate</i>	• Focus on the most dominant first order loop: ○ Assess each pairwise comparison forming the loop using the general GRADE principles ○ Focus on risk of bias, inconsistency, indirectness and publication bias; (not imprecision) ○ Choose the lowest of the certainty ratings • Assess whether intransitivity is present, and rate down further if necessary
Step 3: <i>Rate the network estimate (based on the ratings of the direct and indirect estimates and the assessment of coherence and precision)</i>	• If the direct estimate has high certainty, choose this rating • If only direct or only indirect evidence is available, choose this rating • If both direct and indirect evidence are available: ○ Choose the rating that contributes the most to the network estimate, or choose the highest between the direct and indirect ratings ○ Assess whether incoherence is present, and rate down further if necessary • Assess whether imprecision is present, and rate down further if necessary • Decide which estimate (direct, indirect or network) is preferable as the best estimate of effect

Figure 10 – Flow for rating the certainty of a network estimate



Source: Brignardello 2018⁴⁴

5.6.2 Rating the certainty of the direct (step 1) and indirect (step 2) estimates

Step 1: For rating the certainty of the direct estimates in step 1, the general GRADE principles are applied (cross-reference).

Step 2: For rating the certainty of the indirect estimates in step 2, the GRADE Working Group suggests focusing on (the most dominant) first order loops, if possible. Within this loop, each pairwise (*i.e.* direct) comparison that contributes to the indirect estimate of the comparison of interest is assessed using the general GRADE principles. After doing this, the lowest confidence rating constitutes the confidence rating of the indirect comparison. Next, it is assessed whether intransitivity is present. Intransitivity is present when the indirect comparisons are based on studies with different characteristics that may modify the treatment effect, such as differing patient characteristics, co-interventions, comparators, measurement of outcome, etc. In that case the confidence rating of the indirect evidence would need to be rated down, most certainly if only indirect evidence is available.

The GRADE Working Group introduced two interesting shortcuts for step 1 and 2, making the process more efficient:⁴⁴

Shortcut 1: Consideration of imprecision is typically not necessary when rating the direct and indirect estimates to inform the rating of the network estimates. The reviewer can rate the direct and indirect evidence on the basis of the other four domains, while reserving the rating of precision to the certainty of the network estimate.

Shortcut 2: There is no need to rate the indirect evidence when the certainty of the direct evidence is high, and the contribution of the direct evidence to the network estimate is at least as great as that of the indirect evidence.

5.6.3 Rating the certainty of the network estimates (step 3)

Step 3: If only direct or only indirect evidence is available for a given comparison, the certainty of the network estimate will be based on the estimate of that direct or indirect evidence in step 3. When, for a particular comparison, both direct and indirect evidence are available, an important additional step is to assess the **coherence** between the estimates of this direct and indirect evidence. The GRADE Working Group published a specific guidance on this topic; reviewers evaluating an NMA are strongly advised to read this guidance.⁵³ Coherence refers to the agreement between direct and indirect evidence, and should be dealt with using the following approach:

(A) examining for incoherence by comparing the direction and magnitude of the point estimates, evaluating the extent of overlap of the associated confidence intervals, and considering the results of a statistical comparison of the two estimates;



(B) assess whether both estimates contribute importantly to the network estimate;

(C) explore the cause of incoherence if present, including bias (from limitations in the study design or publication bias), indirectness, or intransitivity. Rating down for incoherence based on bias or indirectness is mostly with one level, starting from the certainty level of the source of evidence with the largest influence on the network estimate, or alternatively the source with the highest certainty. Rating down for incoherence based on intransitivity is not necessary when the network estimate certainty rating is based on the indirect evidence, and the rating of the latter was graded down because of intransitivity, unless imprecision is also present.

The assessment of the precision of the network estimate deserves specific attention, since there are some particularities in the context of NMA.⁵⁴ Three questions – that are very similar to the general GRADE principles – need to be answered:

- (1) Does the confidence interval cross any of the thresholds;
- (2) What is the effect size; and
- (3) Is the optimal information size met?

The principles about how the confidence interval relates to a threshold are the same as those used for the general GRADE assessment: if the confidence interval crosses a threshold, one should rate down for imprecision (up to three levels). If the confidence does not cross a threshold, one should look at the effect size. When the effect is modest (which is a matter of judgement), reviewers should not rate down for imprecision. On the other hand, when the effect size is large, one should assess if the OIS is met. In the context of NMA, because network estimates are calculated using direct and indirect evidence, the approach for determining whether the total sample size meets the OIS often does not work. An alternative approach is to calculate the sample size for the single study that would have generated the network estimate and its confidence interval. The GRADE Working Group provided formulas for these calculations for the relative risk, odds ratio and mean difference.⁵⁴ If this calculated sample size does not meet the OIS, reviewers should rate down for imprecision.

As a final assessment, when reviewers rate down the network estimate for incoherence and/or imprecision, choosing the highest certainty direct or indirect evidence rather than the network estimate as the best estimate of effect may be preferable.

5.6.4 Glossary network meta-analysis and further reading

Table 13 – Network meta-analysis glossary

Concept	Definition
Direct estimate	Estimate of effect provided by a head-to-head comparison (such as trials of A versus B, when A versus B is the comparison of interest)
Indirect estimate	Estimate of effect provided by two or more head-to-head comparisons that share a common comparator (such as trials of A versus C and trials of B versus C, when A versus B is the comparison of interest)
First order loop	Loop that involves only a single additional intervention. For example, if we are interested in A versus B, the direct estimates of A versus C and B versus C constitute a first order loop.
Second order loop	Loop that involves two other interventions, such as A versus C, C versus D, and D versus B, when A versus B is the comparison of interest
Intransitivity	Differences in study characteristics that may modify the treatment effect in the direct comparisons (such as A versus C and B versus C) that form the basis for the



	indirect estimate of effect of the comparison of interest (A versus B), and thus bias the indirect assessment of A versus B. Factors that may modify treatment effects include differing patient characteristics; differing co-interventions; differing extent to which interventions of interest are optimally administered; differing comparators; and differences in measurement of outcome
Incoherence	Differences between direct and indirect estimates of effect

Source: adapted from Puhan et al⁴⁵

ADDITIONAL REFERENCES OF INTEREST:

Brignardello-Petersen R et al.⁵⁵; Brignardello-Petersen R, et al.⁵⁶; Brignardello-Petersen R, et al.⁵⁷; Yepes-Nunez JJ, et al.⁵⁸; Hao Q, et al.⁵⁹

5.7 Summary without meta-analysis (SWIM)

For various reasons, it may not be possible to do a meta-analysis of effect estimates. In that case, other approaches to synthesise the findings of a systematic review and draw conclusions have to be used.

The Cochrane handbook considers the following methods as acceptable for summarising evidence without a meta-analysis:⁶⁰

- Summarising effect estimates
- Combining P values
- Vote counting based on the direction of effect

More details and worked examples of synthesis without meta-analysis can be found in the Cochrane handbook for interventions. The chapter on SWiM contains clear instructions and is as such a recommended read.

5.7.1 Summarising effect estimates

This method is suggested as a possible solution when the estimates of interventions effects are available, but the variances of the effect are not reported or are incorrect (for example in non-standard study designs that involve clustering).

The (standardised) effects can be summarised using descriptive statistics such as the median, interquartile range and the range. Calculating these statistics addresses the question 'What is the range and distribution of observed effects?' Box-and-whisker or bubble plots can complement reporting of the summary statistics by providing a visual display of the distribution of observed effects.⁶⁰

5.7.2 Combining P values

When there is minimal information reported beyond P values and the direction of effect, the types of outcomes and statistical tests differ across studies or non-parametric tests are reported, combining P values can be used as a method to summarise the findings of the studies. Combining P values addresses the question 'is there evidence that there is an effect in at least one study?'⁶⁰

There are several methods available; the Fisher's method is described in the Cochrane handbook. The method used should be reported, alongside the available effect estimates.



5.7.3 *Vote counting based on the direction of effect*

When the direction of effect is reported without further information, or there is no consistent effect measure or data reported across studies, vote counting based on the direction of effect can be considered.

Vote counting is based on comparing the number of studies showing a benefit to the number of studies showing harm for a particular outcome. To categorize studies as showing benefit or harm, only the observed direction of effect is taken into account. Neither statistical significance nor the size of the effect should be taken into account.

Note that vote counting based on statistical significance and vote counting based on subjective rules (based on a combination of direction, statistical significance and /or magnitude of effect) are unacceptable synthesis methods.⁶⁰

An estimate of the proportion of effects favouring the intervention can be calculated ($p = u / n$, where u = number of effects favouring the intervention, and n = number of studies) along with a confidence interval (e.g. using the Wilson or Jeffreys interval methods).⁶¹ Unless there are many studies contributing effects to the analysis, there will be large uncertainty in this estimated proportion.^{60, 61}

It is important to report vote counting as the synthesis method used and to avoid informal use (hidden in sentences such as “The majority of studies show...”).⁶⁰

5.7.4 *Reporting of a synthesis without meta-analysis*

When a method for SWiM is used, the reporting checklist published by Campbell et al (2020) should be applied.⁶²

The following items should be transparently reported:

- Item 1: grouping studies for synthesis
- Item 2: describe the standardised metric and transformation method used
- Item 3: describe the synthesis methods
- Item 4: criteria used to prioritise results for summary and synthesis
- Item 5: investigation of heterogeneity in reported effects
- Item 6: certainty of evidence
- Item 7: data presentation methods
- Item 8: reporting results
- Item 9: limitations of the synthesis

6 GRADE FOR DIAGNOSTIC TESTS AND TEST-TREATMENT MANAGEMENT STRATEGIES

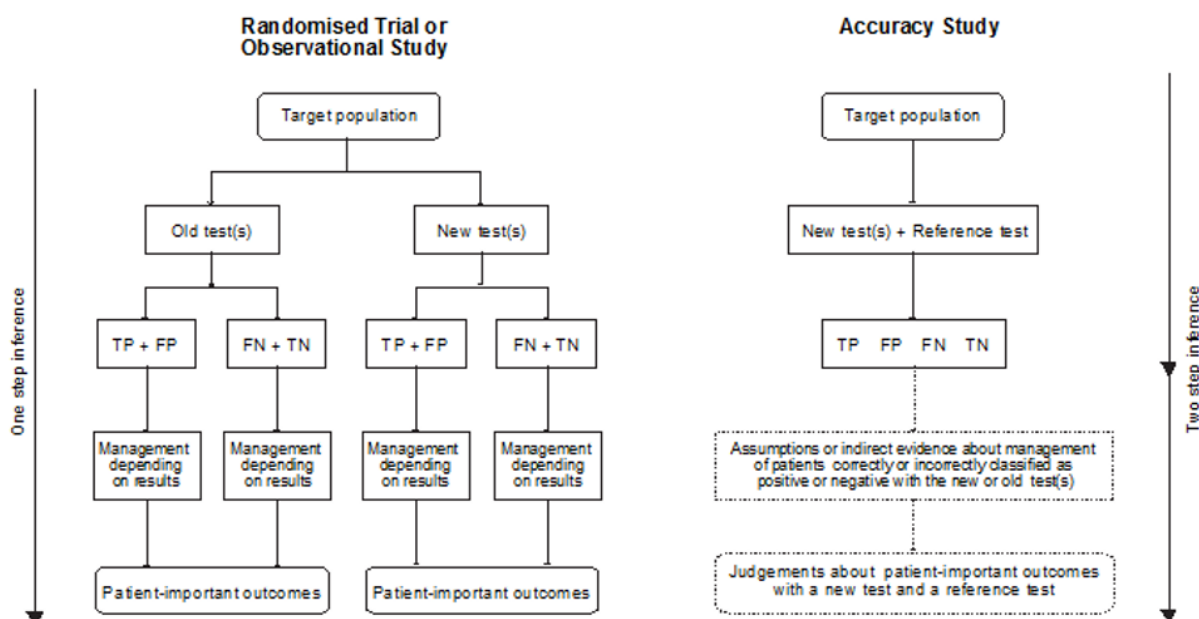
6.1 Introduction (Update 2025)

To apply GRADE and judge the evidence soundly, it is essential that at least one person involved in developing the guideline has profound knowledge of GRADE and is more than able to implement it.

Similar to the GRADE for interventions (see for the general principles of the GRADE steps in 5.2.1), the GRADE framework can be used for diagnostic tests or strategies based on a diagnostic test. The GRADE working group papers guide authors of systematic reviews and guideline developers in assessing the certainty of a body of evidence from DTA studies and strategies.⁶³⁻⁶⁹ The Cochrane handbook offers guidance as well.⁷⁰

Recommendations should consider whether incorporating the new diagnostic test into the treatment strategy will achieve a favourable benefit-to-deficit ratio. This evaluation requires evidence comparing alternative test-treatment strategies focusing on patient-important outcomes. A randomised clinical trial is considered the ideal research design for evaluating the effects of different test-treatment combinations on patient-important outcomes (Figure 11 left side).

Figure 11 – Generic study designs to evaluate the impact of testing



TP: true positive, FP: false positive, TN: true negative, FN: false negative; Source: GRADE Handbook chapter 7⁷¹

For studies designed as basic diagnostic studies (see Figure 11 right side), patient-important outcomes are often not considered or studied. The focus being on test accuracy (sensitivity and specificity), i.e., how well the test classifies patients correctly as having or not having a disease. The underlying assumption is that obtaining a better idea of whether a target condition is present or absent will result in an improved outcome for the patient.

Unfortunately, this assumption might be incorrect. The bottom line is that, if a test or diagnostic strategy fails to improve patient-important outcomes, there is no reason to use the test, regardless of its accuracy. Inferring from data on test accuracy (TA) that a diagnostic test or strategy improves patient-



important outcomes is preferably based on the availability of effective treatment. However, even without an effective treatment, an accurate test may be beneficial if it reduces test-related adverse effects or anxiety, or improves the well-being of patients and caregivers through the prognostic information it provides.

Approach by availability of evidence on patient-important outcomes

- RCT on test-treatment strategies: When diagnostic intervention studies – ideally RCTs (left panel Figure 11) but also non-randomised studies – directly comparing the impact of alternative test-treatment strategies on patient-important outcomes are available, guideline panels can use the GRADE approach as described for intervention studies (see section 5).
- A diagnostic accuracy study:
 - Provides only evidence on the performance of a test. This evidence can be considered sufficient for decision-making in the case of comparative questions with high-certainty evidence about accuracy and potential complications related to the test.
 - If evidence on test accuracy measures is insufficient for decision-making, researchers can make inferences about the likely impact on patient-important outcomes. Evidence or assumptions about the prevalence of the target condition, the direct effects (benefits and harms) of the test, the downstream effects of interventions based on the test results, and the prognosis of patients are needed. This evaluation requires judgements about the relationship between test results and patient-important consequences. For transparency, the ideal approach is to have a fully developed model, with detailed assumptions and calculations. When this is not available, the perceived benefits and harms of the test should provide additional considerations informing the final judgement.

6.2 Framing the question

6.2.1 The purpose of a test or diagnostic strategy

The test of interest or a 'new' test can play different roles in the existing diagnostic pathway

- replace a current or old test if more accurate, less risk involved, less burden to the patient, easier to interpret or quicker results, reduced burdens technically or organisationally, or less costly
- act as a triage to minimise the use of invasive or expensive tests as only patients with a particular result on the triage test continue the testing pathway. The before-test (before the existing diagnostic pathway) can be of benefits due to reduced costs or the simplicity of the test
- act as add-on to enhance accuracy (to limit the number of either false positive or false negative results) of a diagnosis beyond current test.
- parallel or combined new test which is intended to be used concurrently with an existing test. Both, the results of the existing and the parallel test are utilised for making a diagnosis and to determine management

The purpose of a test should be explicit, together with the explanation of its use; the test can include signs and symptoms, imaging, biochemistry, etc. and be used for identifying a physiological derangement or abnormality, establishing prognosis, treatment response, screening or diagnosis.

6.2.2 The format of the question

Applying GRADE for diagnosis begins with formulating an appropriate clinical question using 'PICO' or another structured format.⁶³ The format of the question follows the same principles as the format for intervention questions. The four components are: a defined population (P=population) for whom the test or diagnostic strategy (I=index test) is being considered concerning a comparison test or diagnostic strategy (C=comparison) according to defined patient outcomes of interest (O=outcomes).

Recommendations may differ across subgroups of patients (as may evidence quality). Therefore, different subgroups or settings may need separate questions (and separate evidence summaries).

A test may be compared with no testing, which means that only one test might be mentioned in the research question. Possible formats for diagnostic research questions may be the following:

- Should TEST A vs. TEST B be used in SOME PATIENTS/POPULATION?
- Should TEST A vs. TEST B be used for SOME PURPOSE?

The basic design of an accuracy study has a single group of participants, a single index test and a single reference standard. The Cochrane handbook presents variations on this basic design for each of the three components e.g. multiple reference standards, composite reference standard, a mathematical model as reference standard (latent class analysis), etc.⁷²

Examples from the GRADE book

1: In patients suspected of coronary artery disease (patients) should multi-slice spiral computed tomography (CT) of coronary arteries (intervention) be used as replacement for conventional invasive coronary angiography (comparison) to lower complications with acceptable rates of false negatives associated with coronary events and false positives leading to unnecessary treatment and complications (outcomes)?

This example illustrates one common rationale for a new test – test replacement (coronary CT instead of conventional angiography) to avoid complications associated with a more invasive and expensive alternative for a condition that can effectively be treated. In this situation, the new test would only need to replicate the results of the existing test to demonstrate greater patient net benefit. This assumes that the new test similarly categorises patients at the same stage of the disease and that the consequences of the test result, i.e. management decisions and outcomes, are similar.

Other examples are in the GRADE book:

2: In patients suspected of cow's milk allergy (CMA), should skin prick tests rather than an oral food challenge with cow's milk be used for the diagnosis and management of IgE-mediated CMA.

3: In adults cared for in a non-specialised clinical setting, should serum or plasma cystatin C rather than serum creatinine concentration be used for the diagnosis and management of renal impairment.

When the reference standard, which is the current best and generally accepted test able to distinguish diseased from non-diseased, is imperfect, the estimates of sensitivity and specificity will be biased. There are several methodological options for summarising diagnostic accuracy in the absence of a good reference standard; these aspects go beyond the scope of this GRADE update for diagnosis.



6.3 GRADE DTA Evidence Profiles and Summary of Findings tables (Update 2025)

Evidence is summarised with EP and SoF tables. The SoF table provides the detailed judgements and reasons for grading up or down. As such, an evidence profile provides the information for other guideline developers and validators that allows them to understand how the judgement about the level of evidence was reached. An example for a SoF is shown below.

To enable judgement of the effects of testing, the summary estimates of sensitivity and specificity, and their 95% confidence intervals, are applied to a hypothetical population of – usually – 1000 patients that would undergo the testing because of suspicion of the disease.

To perform the calculations, an estimate of the pre-test probability (prevalence) of the target condition in the population to which the test will be applied is needed.

- If the prevalence is known, e.g. from the literature or registries, it is applied to the 1000-persons GRADE table. Important, the selected prevalence should not be outside the range of the prevalences of the included studies.
- If information is unavailable, the median prevalence of the included studies can be applied. The prevalence estimate, the 25th and 75th percentiles (1st and 3rd quartiles) can be presented to detect variation or imprecision.

Examples of GRADE SoF tables for a prevalence estimate (number of patients with the target condition (N-metastasis) $0.246 \times 1000 = 246$) from the included studies for the following research question: 'Should ultrasound be used for diagnosing nodal metastasis (N-staging) in patients with a primary cutaneous melanoma?'.

Example 1 SoF GRADE table - Effect per 1000 patients tested

Test	No. of studies (No. of patients)	Median prevalence %	Sensitivity (95% CI)	Specificity (95% CI)	Predictive value of positive test (95% CI)	Predictive value of negative test (95% CI)	True positive	False negative	False positive	True negative	GRADE LoE
<i>Ultrasound</i>	13 (5758)	24.6	28.8 (23.3-35.0)	91.6 (89.4-93.5)	53.0 (45.3-60.5)	79.8 (78.43-81.1)	71	175	63	691	Moderate to low

Example 2 GRADE evidence profiles table - Effect per 1000 patients tested

Sensitivity		28.9% (23.3-35.0)							
Specificity		91.7% (89.4-93.5)							
Prevalences		24.6% (22.0-27.4)							
Outcome	No. of studies (No. of patients)	Study design	Factors that may decrease/increase quality of evidence					Effect per 1000 patients tested	Test accuracy QoE
			Risk of bias	Indirectness	Inconsistency	Imprecision	Publication bias	pre-test probability of 24.6%	
True positives (patients N+)	13 studies 1182 patients	Prospective (one large diagnostic study) and retrospective studies. No case-control type studies	low	not serious	very serious ^a	serious ^b	none	71 (57-86)	⊕⊕○○ LOW
False negatives (patients incorrectly classified N-)								175 (160-189)	
True negatives (patients N-)	13 studies 4576 patients		low	not serious	not serious	not serious	none	691 (674-705))	⊕⊕⊕○ Moderate
False positives (patients incorrectly classified as N+)								63 (49-80)	

^a High unexplained heterogeneity of results; ^b wide confidence intervals



6.4 Deciding on the importance of outcomes

6.4.1 Defining outcomes in a GRADE diagnostic framework

GRADE for diagnosis follows the same steps as GRADE for interventions. Ideally, guideline developers will investigate for outcomes on evidence for the test accuracy, direct effects of the test, downstream consequences, and links between test results and management decisions. This can be done by means of systematic reviews. Linking the evidence to the downstream consequences can be done informally (“back of the envelop approach”) or formally in a decision model. If this is not possible, for example because of limited resources, this should be stated.

The GRADE framework structures the outcomes as follows:⁶³

- The patient-important consequences of being correctly classified, usually associated with benefits or a reduction in adverse outcomes; defined by test accuracy measures
 - true positive (TP)
 - true negative (TN)
- The patient-important consequences of being incorrectly classified, usually associated with worse consequences (harms), including failure to treat and failure to potentially reduce burden of disease; defined by test accuracy measures
 - false positive (FP)
 - false negative (FN)
- The consequences of inconclusive results
- The complications of a new index test versus the complications of the reference standard
- The resources used and cost implications

An example of assessing the patient-important consequences of being classified into TP, TN, FP and FN categories can be found below (Table 14).

Table 14 – Patient-important consequences of being classified into TP, TN, FP and FN

<i>Example of linking patient-important outcomes to test results</i>	
Question	In adult patients with suspected obstructive lung disease (COPD), does spirometry testing accurately diagnose COPD, improve clinical outcomes and improve efficiency of care compared to CT-scan?
Population	Adult patients with suspected COPD
Index test	Spirometry
Comparator	CT
Outcomes	Consequences
TP	Effective treatment and improved QoL Side-effects of therapy Improvement in comorbid conditions (e.g., hypertension) Reduced risk of CV events, post-CV events, lung infection, overall health costs
TN	Confirmation of absence of COPD; Psychological relief Possible repeat testing if patient deemed at high risk for COPD Consideration of alternative causes for symptoms; Focused treatment on true cause of symptoms Saves time and resources
FP	Unnecessary treatment and utilisation of resources; Increased costs due to treatment; Side-effects of therapy Psychological distress; Delay in diagnosis of true condition
FN	Absence of necessary treatment; Reduced QOL; Psychological distress

	Possible repeat testing if patient deemed at high risk for COPD Risk of lung infection, hypertension, CV events, Post-MI events, Post-stroke events, Death Increased costs and utilisation of resources due to other condition(s)
Inconclusive results	Proportion of patients in whom spirometry doesn't provide a result
Complications of the test	Index: Hyperventilation; Comparator: Ionisation
Costs	Not addressed

True positive (TP), false positive (FP), false negative (FN), true negative (TN)

When assessing a test for screening purposes, the analytic framework developed by the U.S. Preventive Services Task Force (USPSTF) can be consulted.⁷³

6.4.2 Outcome-specific decision thresholds (Update 2025)

For DTA projects, GRADE recommends setting threshold(s) when assessing the accuracy of a single test when no direct evidence on patient-important outcomes is available.⁷⁴ The threshold(s) are set taking into consideration consequences of testing and effects on patient-important outcomes (see 6.4.1) and should not be confused with the positivity threshold or cut-off value. A common method is to define 3 thresholds to arrive at very accurate (VA), accurate (A), inaccurate (I), and very inaccurate (VI) intervals for the test under study. This works best when applying the EtD framework in the next phase, but two thresholds (A and I) might be preferred e.g. in SR or HTA, when the focus is on judging accurate versus inaccurate.

The definition for threshold(s): the acceptable number or proportion of true and false negative and positive results to support conclusions for systematic reviews and health care decisions for guidelines.

The GRADE guidance 37 describes how modelling, informal and formal, is used to set thresholds.⁷⁴ The formal way needs expertise and will be difficult to implement hence not recommended at KCE. An informal method starts with developing the analytical framework (see 6.4.1) and the consequences for patients of a FN and FP result.

Example

Consider a systematic review of diagnostic accuracy of serologic testing of total antibodies for an infectious disease e.g. COVID-19 in nursing home residents. After considering the consequences of the test results, a threshold of 50 out of 1000 for FN (A||) might be chosen by the experts of the review or GDG (below 50 A, above 50 I). A meta-analysis result of 150 (95% CI: 110-210) would be considered an invalid (I) test.

Another informal way was proposed by Korevaar et al. To establish a threshold, they propose to ask the question: "Considering x number of people suspected of the target condition, how many FN and FP results are acceptable, considering the potential harms of such misclassifications?"⁷⁵ Korevaar et al. stresses that minimally acceptable criteria for test performance should form the basis for hypothesis formulation and sample size calculations in diagnostic accuracy studies.



6.5 Presenting and visualising DTA results (Update 2025)

Presenting diagnostic test accuracy results in a systematic review requires approaches that convey both accuracy measures (sensitivity, specificity, predictive values, likelihood ratios) and the uncertainty around them. Here's an overview of the several possibilities.

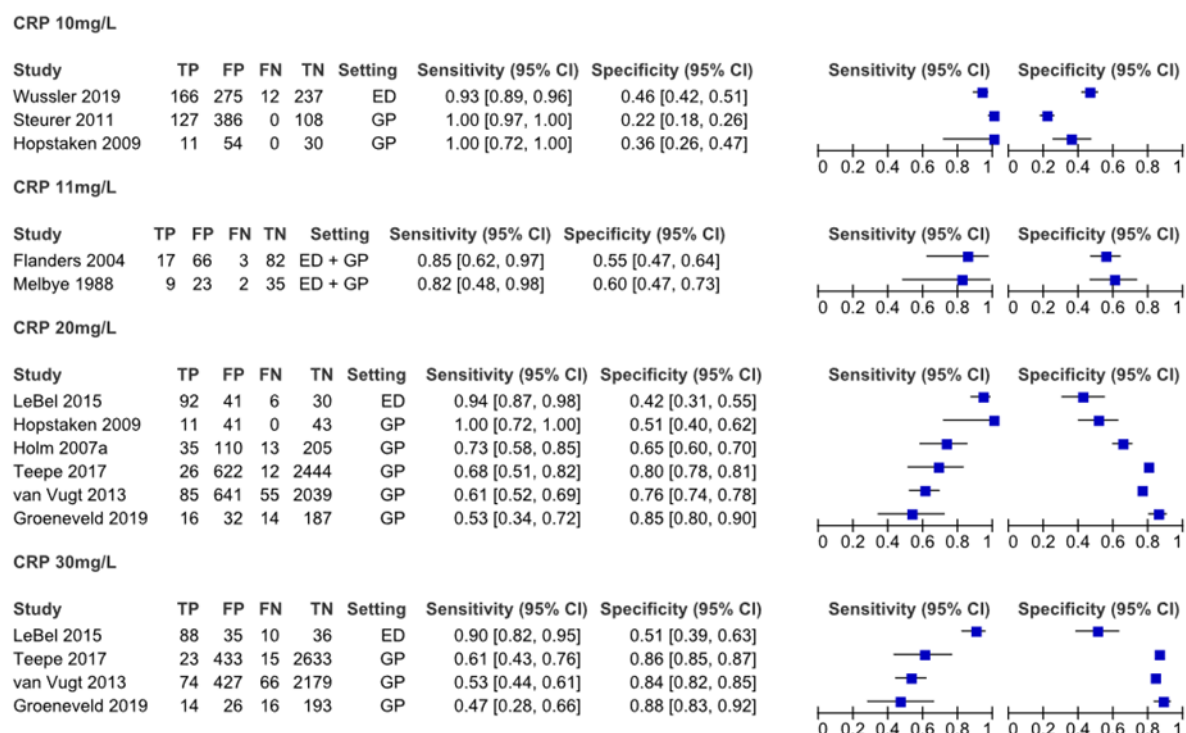
6.5.1 Text and Summary Tables

- Basic 2×2 table per study (TP, FP, TN, FN) could be handy and very transparent, but quickly become too large when the number of studies increase.
- Summary tables listing sensitivity, specificity, and 95% confidence intervals (95%CI) for each study help readers to see study-level results and variability. Again, these tables may become too large. The results are more often incorporated in Forest Plots and Receiver Operating Characteristic (ROC) plots (see below).
- Summary tables can however be useful when presenting the results of a bivariate meta-analysis (see Section 6.6.2).

6.5.2 Forest Plots

- Sensitivity and specificity for each study are plotted side-by-side (=mirrored/coupled plots): each study is represented with its point estimates and 95%CI.
- Useful to show the range of estimates across studies, and whether results are consistent.
- Allows visual comparison of trade-offs across studies. Typically ordered by descending sensitivity, ascending threshold or study size.

Figure 12 – Example Forest plot of C-reactive protein (CRP) for the diagnosis of pneumonia, ordered by: CRP threshold, setting, and sensitivity



Generated with RevMan; Source: Struyf T et al.⁷⁶; ED: emergency department, GP: General practice setting

HOW TO INTERPRET THIS FOREST PLOT?

- Clear threshold effect of CRP: the higher the CRP threshold, the lower the sensitivity and the higher the specificity.
- Setting modifies test performance: CRP is less sensitive in general practice, probably because of the spectrum of patients seen in that setting (patients are earlier in disease course, with (on average) less severe presentations).
- Inverse correlation between sensitivity and specificity across all thresholds: as sensitivity falls, specificity rises. This is an expected clinical trade-off in diagnostic testing, particularly for continuous biomarkers like CRP:
 - Low thresholds should be used for ruling out a disease, as they prioritise sensitivity (fewer false negatives). This is useful when the goal is to safely exclude a serious condition.
 - High thresholds should be used for ruling in a disease, as they prioritise specificity (fewer false positives). This is useful when the goal is to confirm a diagnosis and avoid unnecessary treatment.
- Confidence intervals are wide in smaller general practice studies, reflecting lower sample sizes. Narrow in large emergency department studies, adding precision.
- The direction of the threshold effect is very consistent across studies.
- There is clear heterogeneity across studies, both within the same threshold and across thresholds. Much of it is explained by threshold effects (as the CRP cut-off increases, sensitivity decreases and specificity increases). Additional heterogeneity is likely due to setting differences (e.g. GP versus ED) and study-level factors. To disentangle this more clearly, plotting these same data in a ROC space (sensitivity versus 1-specificity, see 6.5.3) would show whether the variability aligns along a curve (threshold-driven) or is scattered (true inconsistency).
- Clinical usefulness: Lower thresholds (10 to 20 mg/L) could be good for “rule out” strategies in ED or GP. Higher thresholds are much better as “rule in” tests, particularly in higher prevalence settings.
- Implication for practice: there is no single best threshold. The choice depends on clinical aim (rule out versus rule in) and setting (primary care versus emergency).

The way the data is presented should align with the review's primary objective. If the review aims to identify "red flags" for a disease, specificity is likely more important, and the forest plot might be ordered by descending specificity. Conversely, if the review seeks to find tests to safely rule out a disease, sensitivity should be prioritised in the presentation.

When there are important covariates in the review that are likely to influence the accuracy of the test(s), such as age or setting, order the Forest Plot results according to these covariates.

When assessing trade-off between sensitivity and specificity, ROC plots may be a more interesting figure than a forest plot.

WHAT TYPES OF DATA DO WE NEED TO CONSTRUCT A FOREST PLOT?

- Raw 2x2 data: true positives (TP), false positives (FP), false negatives (FN), true negatives (TN)
- If not available, we need at minimum the sensitivity and specificity of an index test (at defined thresholds), in combination with the prevalence of the target disease in the study population, in order to construct the 2x2 table



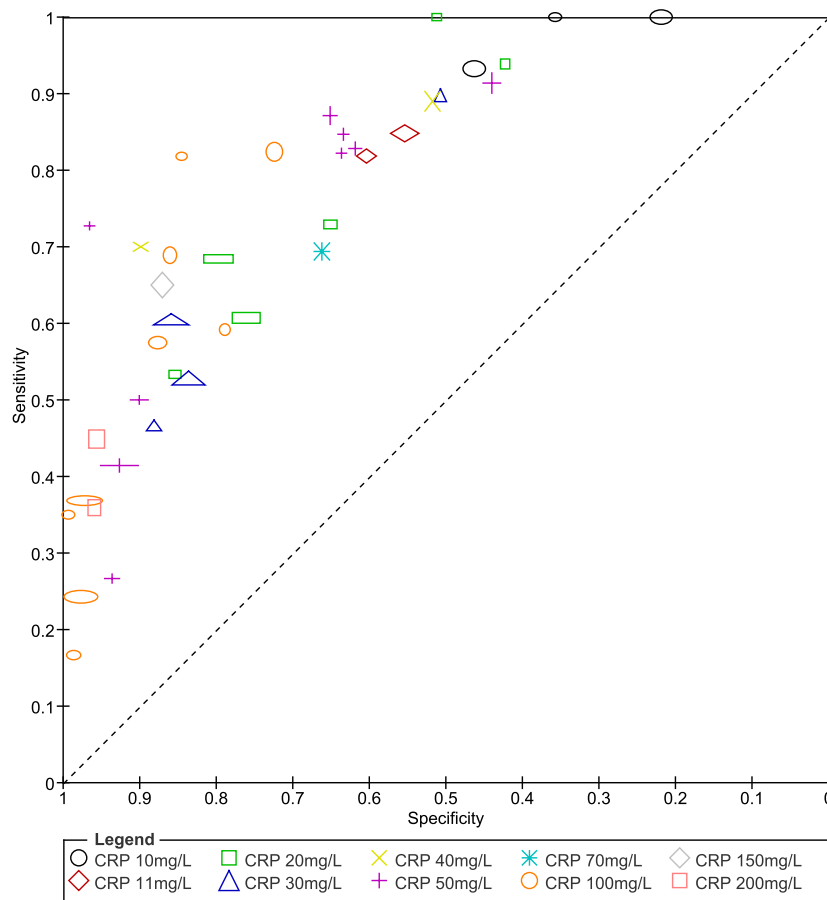
6.5.3 Receiver Operating Characteristic (ROC) plots

ROC plots show sensitivity versus 1–specificity. Each study is a point, and the size of the point/symbol may reflect study sample size.

An ROC plot is more interesting than a Forest Plot, when:

- Threshold effects are important: If studies use *different cut-offs* for the same biomarker (like CRP), the trade-off between sensitivity and specificity is key. An ROC curve shows this relationship directly, whereas a forest plot lists sensitivity and specificity separately.
- You want to show the overall pattern of accuracy: ROC space allows you to visualise whether a test tends to operate in the “good” zone (upper-left corner = high sensitivity *and* specificity).
- You want to assess heterogeneity across studies. ROC plots can distinguish between threshold heterogeneity (the studies align along a curve), or true inconsistency in accuracy (the studies are scattered widely off the curve). Forest plots, by treating sensitivity and specificity separately, cannot show this distinction. Forest plots fragment this information and can give a misleading impression of heterogeneity. As a result, Forest Plots can exaggerate apparent heterogeneity (since one metric may look highly variable while the other looks stable).
- Comparing two (or more) tests: ROC plots are excellent for showing if Test A’s operating points cluster closer to the “ideal” corner than Test B’s. Especially useful in reviews where you want to highlight relative accuracy.
- Summarising meta-analysis results: Summary ROC (SROC) curves from hierarchical models (bivariate or HSROC) give a pooled operating point or a curve respectively, and a confidence and/or prediction region. This illustrates not just pooled accuracy, but how results might vary in future studies. The presentation of meta-analyses is discussed in section 6.6.
- Visualising inverse correlation: ROC plots make it obvious that sensitivity and specificity are inversely related. Forest plots list them separately and require side-by-side interpretation.

Figure 13 – ROC plot of C-reactive protein (CRP) for the diagnosis of pneumonia: same data as the Forest Plot example



Generated with RevMan; Source: Struyf T et al.⁷⁶

While adding a covariate to an ROC plot is possible, the plot can become more chaotic and difficult to interpret. Ordering by a covariate is more recommendable when using Forest Plots or when presenting results of a meta-analysis (see 6.6.2).

WHAT TYPE OF DATA DO WE NEED TO CONSTRUCT A ROC PLOT?

- Raw 2x2 data (TP, FP, FN, TN)
- If not available, we need at minimum the sensitivity and specificity of an index test (at defined thresholds), in combination with the prevalence of the target disease in the study population. The total sample size of the study is also needed if we want to scale the size of the symbols in the ROC plot according to the study size.



6.5.4 Dumbbell plots

A Dumbbell plot uses three dots connected by lines (Figure 14):

- Blue dot: probability of disease *before testing* (pre-test probability, often prevalence or baseline risk).
- Green dot with a minus sign (–): probability of disease *after a negative test result* (post-test probability given a negative test result = 1 *minus* Negative Predictive Value (1-NPV)).
- Red dot with a plus sign (+): probability of disease *after a positive test result* (post-test probability given a positive test result = Positive Predictive Value (PPV)).

The green and red dots are connected to the blue one, which visually shows how test results shift the probability of a disease. Of course, these colours are flexible and can be adapted, for example, to improve readability for the colour-blind. To this end, you could use a mid-toned blue dot for the pre-test probability, a bright orange for a positive test result, and a lighter blue or a dark grey for a negative one. You could also consider altering the shapes of the dots.

Dumbbell plots summarise the impact of a diagnostic test in terms of individual-level probability shifts:

- They show how much the test result changes certainty compared to the pre-test probability.
- If the green dot is close to the blue one, this means that a negative result doesn't rule out much (poor ruling-out power).
- If the red dot is far from the blue one, this means that a positive result strongly increases certainty (good ruling-in power).
- If both dots are far from the blue dot in opposite directions, this means that the test meaningfully separates sick versus healthy patients.

They make it easy to see which tests provide clinically relevant changes, and whether a test is more useful for ruling in (+) or ruling out (–).

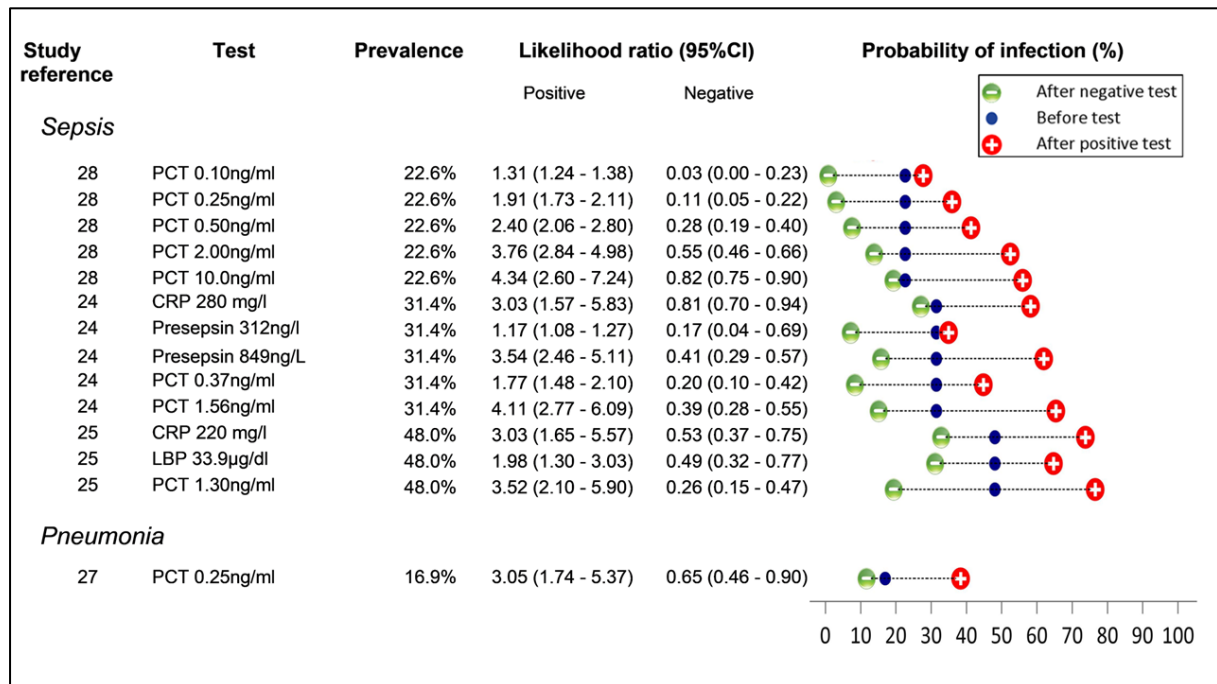
In DTA systematic reviews, where many tests are compared, Dumbbell plots help summarise across studies or populations. They are interesting to use when you want to communicate test performance in probability terms (risk communication), which is easier for clinicians and policymakers than sensitivity/specificity.

Cave!

- Post-test probabilities depend on the prevalence of the target disease or assumed pre-test probability. If this isn't representative, the plot can mislead. If you only have sensitivity and specificity, you can't directly get predictive values without assuming a prevalence.
- The plot ignores uncertainty, because it shows point estimates only, without 95%CI.
- The plot reduces accuracy to just probability shifts, ignoring nuances like heterogeneity across populations, spectrum effects, or study quality.
- The interpretation of the plots requires familiarity with Bayesian thinking and post-test probability. It is not always intuitive for readers used to sensitivity/specificity, although it aligns well with clinical reasoning.

Because of these limitations, it is good practice to combine the plot with a table showing the positive and negative likelihood ratios (LR+ and LR–) of the tests and their 95%CI, as shown in Figure 14. The Likelihood Ratio (LR) is the likelihood that a given test result would be expected in a patient with the target disorder compared to the likelihood that that same result would be expected in a patient without the target disorder.

Figure 14 – Dumbbell plot of blood tests for the diagnosis of sepsis and pneumonia in older adults visiting emergency care.



Source: Struyf T et al.⁷⁶. *The Likelihood Ratio (LR) is the likelihood that a given test result would be expected in a patient with the target disorder compared to the likelihood that that same result would be expected in a patient without the target disorder.

WHAT CAN WE LEARN FROM THIS FIGURE?

Magnitude of the shift matters:

- Tests that move the green dot far *left* (lower probability after a negative test) are good at ruling out disease.
- Tests that move the red dot far *right* (higher probability after a positive) are good at ruling in disease.
- If green/red dots are close to the blue one, the test has little diagnostic value.

Comparison of different thresholds:

- Example: For PCT 0.10 ng/ml (prevalence 22.6%), the green dot is very low (~0–5%), showing a very low post-test probability after a negative → strong for ruling out.
- In contrast, for PCT 10.0 ng/ml at the same prevalence, the negative test barely lowers probability (~15–20%), but the positive test strongly increases the probability (~60–70%) → stronger for ruling in.



Dependence on prevalence

- Look at the bottom group (48% prevalence, e.g. CRP 220 mg/l): even after a negative result, the post-test probability stays ~20–30% → the test can't confidently rule out disease in higher-prevalence populations.
- So: the same test behaves differently depending on baseline risk!

Visual trade-offs

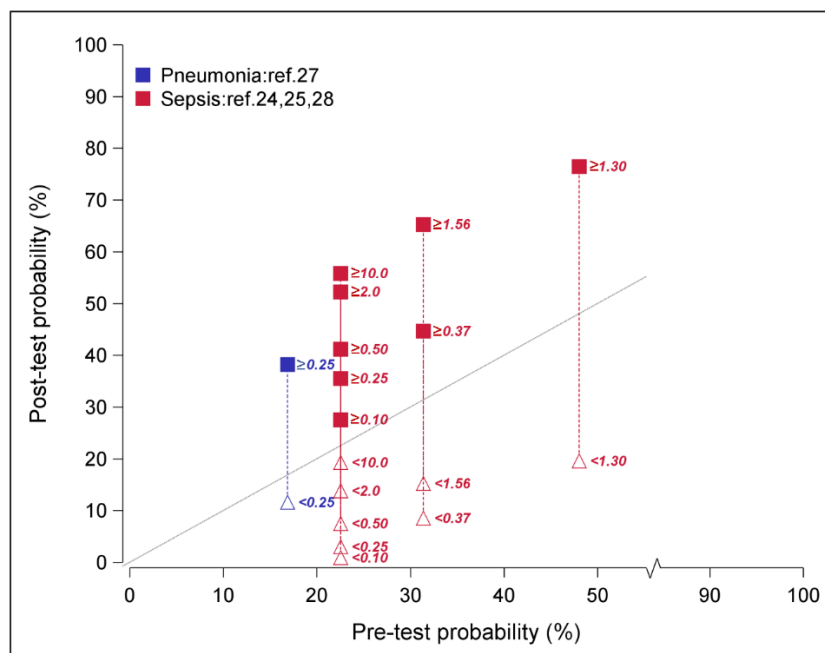
- Some tests give large shifts both ways (e.g. PCT 0.25 ng/ml, pneumonia subgroup: from ~17% → ~4% if negative, or → ~40–50% if positive).
 - Others are asymmetric, mainly useful in one direction (ruling in versus ruling out).
- Dumbbell plots reveal clinical utility:
- Which tests are only marginally useful (green/red dots on top of the blue one).
 - Which tests can truly change management (dots move probability far enough to cross clinical decision thresholds).
 - Low LR– tests (e.g. PCT at low cut-offs) can effectively exclude disease in lower-prevalence settings.
 - High LR+ tests (e.g. PCT at higher cut-offs, CRP) are more helpful for confirming disease.

6.5.5 Probability plot

WHAT ARE PROBABILITY PLOTS (ALSO CALLED PROBABILITY-MODIFYING PLOTS)?

- The X-axis shows the pre-test probability (%): the estimated probability that a patient has the condition *before* testing (often the prevalence in the study population)
- The Y-axis shows the post-test probability (%): the probability of disease *after* knowing the index test result.
- Symbols are used to indicate positive test results and negative test results. These symbols are often connected by a (dashed) line, showing the movement from pre-test to post-test probability for each index test threshold.
- The diagonal line (45°) should be interpreted as a “no information” line: if a test had no diagnostic value, post-test probability would equal pre-test probability. The further away from this line, the more impact the test has (cfr. ROC plots in 6.5.3).

Figure 15 – Probability plot for the use of procalcitonin at different cut-offs for the diagnosis of sepsis or pneumonia in older adults visiting emergency care



Source: Struyf T et al.⁷⁶

WHEN TO USE THIS PLOT?

- Communicating test impact: Clinicians often find it easier to interpret changes in probabilities (risk before and after testing) rather than abstract measures like sensitivity, specificity, or LRs.
- Visualising threshold effects: It shows how test results influence clinical decision thresholds (e.g. at what pre-test probability + at which index test cut-off value would a test meaningfully change management).
- Comparing conditions/tests: You can show at a glance how much a test improves (or fails to improve) diagnostic certainty across diseases or cutoffs.
- Bridging evidence to practice: Helps guideline developers and clinicians see whether a test is useful in real-world decision making.
- Probability plots are most valuable in DTA reviews when dealing with continuous or threshold-based tests (like lab markers), where you can show how different cut-offs change post-test probability. For binary tests (symptoms, yes/no findings), they can be used, but they usually don't add much information.

WHEN IS IT LESS USEFUL?

- Not good for meta-analytic synthesis: It's descriptive and illustrative but doesn't summarise heterogeneity or provide pooled estimates.
- Assumes LRs are stable: These plots are only as reliable as the likelihood ratios used; if those come from biased or heterogeneous studies, the plot may mislead.
- Not suited for rare or extreme pre-test probabilities: If prevalence is very low or very high, shifts in probability may look dramatic but could still be clinically trivial.
- No measure of uncertainty: Unless you add CIs, readers may overinterpret precision.



6.5.6 When to use which plot?

Table 15 – When to use which plot for presenting DTA results: overview

Plot type	What it shows	When to use it	Strengths	Limitations
Text & Summary tables	2×2 data per study, or study-level sensitivity/specificity with 95%CI	Early in results section; small numbers of studies; to give transparency on study-level accuracy	Transparent, easy to check calculations, allows tabular overview of study characteristics	Quickly becomes too large with many studies; not visual; limited synthesis
Forest plots	Sensitivity and specificity (with 95%CI) plotted per study, side by side	When you want to show range of estimates across studies; highlight heterogeneity; order by sensitivity, specificity, threshold, covariate, or sample size	Clear display of study-level variation; allows ordering to show threshold or covariate effects	Separates sensitivity & specificity (harder to see trade-offs); may exaggerate heterogeneity; less good for threshold effects
ROC plots (study-level)	Sensitivity vs 1–specificity; each study a point (size ~ study weight); diagonal line = no diagnostic value	Best when threshold effects are present; for comparing two or more tests; when visualising overall pattern (accuracy zone) or heterogeneity	Shows trade-off directly; reveals if variability is threshold-driven vs inconsistent; excellent for comparing tests	Harder to incorporate covariates (can become messy); less intuitive for some clinicians
Summary ROC (SROC) curves or summary points	Meta-analytic pooled ROC curve or summary points with 95%CI/prediction region; diagonal line = no diagnostic value	After bivariate/HSROC meta-analysis; to summarise accuracy & heterogeneity; when review focus is overall test performance	Shows pooled performance, uncertainty, expected future variability	Technical to produce; needs hierarchical modelling; less intuitive for clinicians
Dumbbell plots	Pre-test probability (blue), post-test after negative (green), post-test after positive (red)	When you want to show impact of test results in probability terms; excellent for guideline discussions; useful for comparing across tests or thresholds	Intuitive for clinicians/policymakers (risk communication); highlights ruling in vs ruling out power	Ignores uncertainty (need for additional table); depends on assumed prevalence; oversimplifies heterogeneity
Probability plots (probability-modifying plots)	Pre-test (x-axis) vs post-test probability (y-axis); symbols show thresholds; diagonal line = no diagnostic value	Best for continuous or threshold-based tests (e.g. lab cut-offs); when demonstrating threshold effects; to show how results shift clinical decision thresholds	Clinically intuitive; directly links test to management; good for comparing cut-offs/tests	Descriptive only (not meta-analytic); assumes stable LRs; no uncertainty unless added; less useful for simple binary tests

Use Forest plots when you want to show study-level estimates and covariate effects; Use ROC plots (study-level or SROC) when threshold effects and accuracy patterns matter; Use Dumbbell plots or Probability plots when you want to translate statistical accuracy into clinically meaningful probability shifts, especially for guideline or practice-oriented communication. Combine these plots with tabular information on likelihood ratios (+95%CI); Use tables for transparency and detailed reporting, but combine them with visual plots for synthesis. Only use them when the number of studies are small.

6.6 Meta-analysis of DTA studies (Update 2025)

6.6.1 Introduction

Meta-analysis in DTA reviews combines results from primary studies evaluating the same index test, in order to provide more precise estimates of test performance and to explore sources of variability. Unlike intervention reviews, which usually summarise a single effect measure such as a risk ratio, DTA reviews must account for the paired nature of outcomes (sensitivity and specificity) which are correlated and can vary according to study design, populations, and test thresholds.⁷⁰

When studies report complete 2×2 data (TP, FP, FN, TN), the recommended approach is to apply hierarchical models that appropriately account for within- and between-study variability. Two models are widely used: the bivariate random-effects model, which directly models sensitivity and specificity as correlated measures and estimates a *summary point* in the ROC space, and the hierarchical summary receiver operating characteristic (HSROC) model, which focuses on estimating and summarising the *ROC curve*.^{77, 78} These approaches allow for the calculation of summary points (pooled sensitivity and specificity), summary curves, and the investigation of heterogeneity through covariates.

The decision to conduct a meta-analysis in a DTA review requires careful judgement. Meta-analysis may be inappropriate when there is substantial clinical or methodological heterogeneity, for example, due to differences in index test execution, thresholds, reference standards, or populations studied. Moreover, hierarchical models generally require a minimum of four studies, and preferably a larger number, to provide stable parameter estimates.⁷⁰ **In situations where data are sparse or too heterogeneous, reviewers are encouraged only to present results descriptively using forest plots and ROC plots of individual study results, without attempting statistical pooling.**

6.6.2 Bivariate random-effects meta-analysis

The bivariate random-effects model is the recommended approach in DTA meta-analyses when studies provide 2×2 data at a similar threshold. It is a random-effects meta-analysis model that estimates sensitivity and specificity jointly, allowing for both within-study correlation between sensitivity and specificity, and between-study heterogeneity in both measures.

It models sensitivity and specificity simultaneously (hence “bivariate”), accounts for their correlation, and allows exploration of heterogeneity. This makes it more appropriate than separate univariate meta-analyses of sensitivity and specificity, which ignore their statistical dependence.

If full 2×2 data aren’t available, reported sensitivity and specificity with confidence intervals can be used to calculate the 2×2 data. At least 4 studies with ‘acceptable’ heterogeneity are needed (ideally ≥10) to get stable variance estimates. We can use bivariate meta-analysis when studies report results at a single common threshold (or when thresholds are reasonably comparable, from a clinical point of view) and when the review aim is to provide a summary point (pooled sensitivity AND specificity) and to explore heterogeneity (via covariates).

If studies report diagnostic accuracies at different index test thresholds, you should use the HSROC model instead (see section 6.6.3). This model summarises the entire ROC curve rather than a single point, which also accounts for the inherent correlation of multiple results from single studies that report on several thresholds (Steinhauser method, see section on HSROC).⁷⁹

The results of a bivariate meta-analysis can be presented in tables (cf. Table 16) or in a figure (cf. Figure 16). A table is most useful when you want to present exact numerical results:

- Summary point estimates: pooled sensitivity and specificity with 95% confidence intervals.
- Heterogeneity and model parameters: between-study variance estimates (τ^2), correlation coefficients, or diagnostic odds ratios.

Remark: τ^2 estimates are rarely shown because they don’t help most readers understand variability in accuracy. While statistically important, correlation coefficients are not straightforward to interpret clinically (e.g. a correlation of –0.6 doesn’t translate into an obvious “real-world” message). Finally,



DORs are mathematically elegant but clinically unhelpful because it masks the trade-off between sensitivity and specificity.

- Subgroup or covariate analyses: showing how sensitivity/specificity changes across different test types, settings, or thresholds. Figures, such as ROC plots with subgroup overlays and confidence regions make it easier to grasp the magnitude and direction of subgroup or covariate effects.

Tables are best for precision and detail, allowing readers to extract the exact numbers for decision-making or secondary analyses. Most Cochrane DTA reviews use both: a table for precision and a figure for intuition.

Figures are particularly valuable to visualise patterns, variability, and uncertainty:

- Summary ROC (sROC) plots: showing each study as a point (sensitivity versus 1–specificity), with the summary curve and confidence/prediction regions. This illustrates heterogeneity and threshold effects.
- Forest plots of sensitivity and specificity: displaying study-level estimates with confidence intervals, alongside the pooled estimate. This helps readers assess variation at a glance.
- Prediction regions from the bivariate model: giving a sense of the range of test performance expected in future studies.

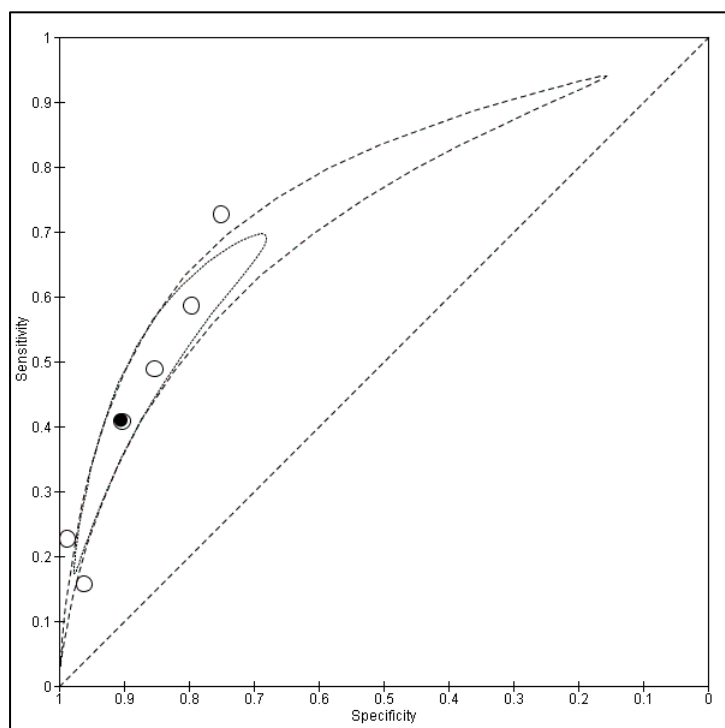
Table 16 – Summary point statistics of selected index tests from a bivariate meta-analysis

<i>Index test</i>	<i>Number of studies (N)</i>	<i>Number of Covid-19 positives/ Total number of participants n/N (%)</i>	<i>Sensitivity (%, 95%CI)</i>	<i>Specificity (%,95%CI)</i>	<i>LR⁺¹ (95%CI)</i>	<i>LR⁻² (95%CI)</i>	<i>DOR³ (95%CI)</i>
All cross-sectional studies							
Cough	25	3207/15459 (20.7%)	67.4% (59.8% to 74.1%)	35.0% (28.7% to 41.9%)	1.036 (0.969 to 1.107)	0.933 (0.816 to 1.067)	1.110 (0.909 to 1.356)
Anosmia	11	2305/9552 (24.1%)	28.0% (17.7% to 41.3%)	93.4% (88.3% to 96.4%)	4.254 (3.172 to 5.705)	0.771 (0.676 to 0.879)	5.549 (4.089 to 7.532)
Ageusia	6	1893/7393 (25.6%)	24.8% (12.4% to 43.5%)	91.4% (81.3% to 96.3%)	2.876 (2.021 to 4.092)	0.823 (0.712 to 0.951)	3.495 (2.408 to 5.072)
Anosmia or ageusia	6	1589/8142 (19.5%)	41.0% (27.0% to 56.6%)	90.5% (81.2% to 95.4%)	4.306 (3.002 to 6.177)	0.652 (0.542 to 0.785)	6.602 (5.271 to 8.270)
Sensitivity analysis: cross-sectional studies with a prospective data-collection only							
Cough	7	1484/6411 (23.1%)	66.3% (57.8% to 73.8%)	40.7% (33.6% to 48.3%)	1.118 (1.005 to 1.243)	0.829 (0.686 to 1.001)	1.349 (1.008 to 1.805)
Anosmia	8	2129/8518 (25.0%)	29.1% (18.9% to 42.1%)	92.3% (85.8% to 95.9%)	3.765 (2.783 to 5.092)	0.768 (0.682 to 0.866)	4.900 (3.717 to 6.460)
Ageusia	5	1843/7293 (25.3%)	29.4% (15.1% to 49.5%)	89.0% (77.6% to 94.9%)	2.667 (1.957 to 3.636)	0.793 (0.669 to 0.941)	3.362 (2.382 to 4.746)
Anosmia or ageusia	5	1534/7406 (20.7%)	36.5% (24.0% to 51.2%)	92.4% (84.1% to 96.5%)	4.782 (3.182 to 7.185)	0.687 (0.586 to 0.806)	6.955 (5.195 to 9.312)

¹LR+= positive Likelihood Ratio; ²LR-= negative Likelihood Ratio; ³DOR= Diagnostic Odds Ratio; Source: Excerpt from a table for illustrative purposes from Struyf T et al. Cochrane Database of Systematic Reviews⁸⁰



Figure 16 – Summary ROC plot for the presence of anosmia or ageusia for the diagnosis of COVID-19 (6 studies)



Source: Struyf T et al.⁸⁰

How to interpret this plot:

- Each open circle represents one included study. The spread of circles shows heterogeneity in accuracy: some studies report higher sensitivity, others lower. In this case, sensitivity is generally low to moderate (around 20% to 70%), while specificity is moderate to high (around 80% to 95%). This suggests the symptom is better at ruling in COVID-19 (when present), than at ruling it out (when absent).
- The solid black circle represents the pooled estimate of sensitivity and specificity from the bivariate model (= summary point estimate), indicating a summary sensitivity around 40%, and a summary specificity of around 90%.
- The smaller dashed ellipse around the summary point indicates the 95% confidence region. It indicates the uncertainty in the pooled estimate, i.e. where the “true” average accuracy lies across these studies. Here it is relatively tight, showing the summary estimate is fairly precise given the available data.
- The wider dashed ellipse is the 95% prediction region. It shows where we would expect results of a new study to fall, with 95% probability. This is wide, showing considerable between-study heterogeneity. It indicates that the accuracy of ‘anosmia or ageusia’ for diagnosing COVID-19 may vary substantially across settings and populations.

Several factors are likely to explain this heterogeneity. First, differences in study populations and settings (hospital vs. community or primary care) influence the spectrum of disease and the prevalence of alternative causes of smell or taste disturbance. Second, variability in how symptoms were recorded (self-report questionnaires versus clinician-led interviews) may have introduced methodological inconsistency. Third, differences in the reference standards used to confirm COVID-19 status could contribute to misclassification, although most studies used PCR testing. Finally, although anosmia and ageusia are conceptually binary symptoms, variation in the definitions applied across studies (e.g.

complete versus partial loss, or 'any alteration' in smell/taste) may have acted as implicit threshold effects, leading to systematic differences in sensitivity and specificity.

While differences in population and study methods may contribute, the shape of the confidence region suggests threshold effects are the most likely explanation, with broader symptom definitions yielding higher sensitivity but lower specificity. Clinicians should therefore interpret pooled estimates cautiously, as the diagnostic value of these symptoms depends strongly on how they are defined and assessed.

6.6.3 Hierarchical Summary ROC (HSROC) Curves

The Hierarchical Summary Receiver Operating Characteristic (HSROC) model is a statistical approach used to synthesise diagnostic test accuracy data when studies evaluate a test at different thresholds. Unlike the bivariate model, which focuses on estimating pooled sensitivity and specificity at a *common* threshold, the HSROC model accommodates *variation in thresholds* across and within studies by summarising accuracy as a curve rather than a single point estimate.

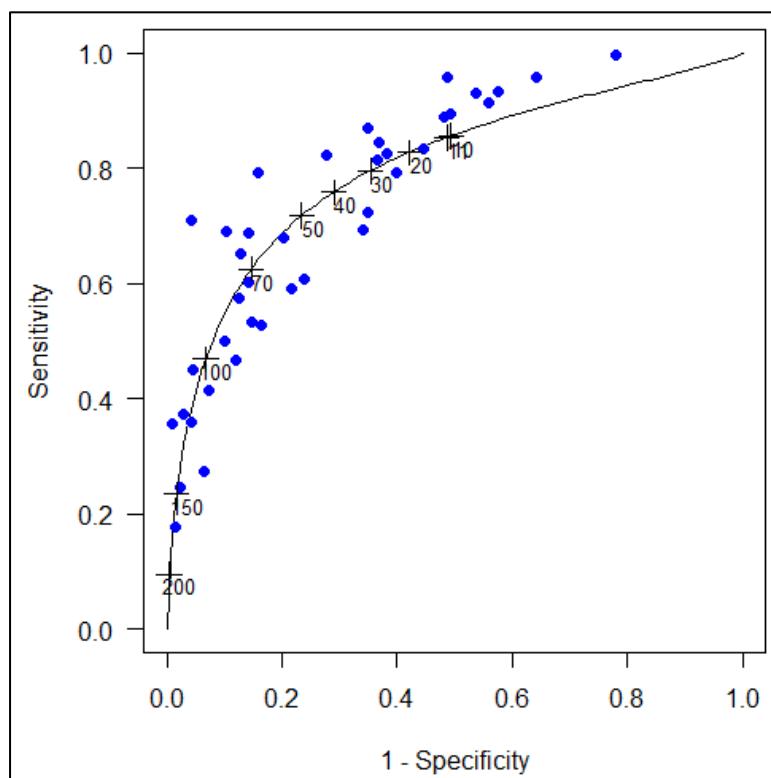
The HSROC curve represents the trade-off between sensitivity and specificity across thresholds, providing a visual and statistical summary of overall test performance. You should use the HSROC approach when:

- Included studies use different thresholds.
- Threshold effects are suspected to be a major contributor to heterogeneity.
- The aim is to provide an overall picture of test performance rather than pooled estimates at one threshold.

When presenting HSROC results, reviewers should include the summary curve with confidence and prediction regions (see Figure 17). This allows readers to see both the average accuracy across thresholds and the expected range of results in future studies. As recommended by the Cochrane Handbook, interpretation should emphasise the implications of threshold effects, sources of heterogeneity, and how these may impact clinical decision-making.



Figure 17 – HSROC analysis of C-reactive Protein for the diagnosis of pneumonia in adults in ambulatory care



Blue dots indicate individual study results, but some studies have multiple blue dots at different CRP thresholds. To make the plot clearer for illustrative purposes, the prediction region was omitted. Source: Biomarkers for serious infections in adults: a systematic review of diagnostic accuracy studies. (in preparation for publication). This plot was produced using the *Diagmeta* package in R (See section e for more information on analysis tools).

Interpretation of the HSROC figure

Each blue dot represents an individual study, plotting its sensitivity against 1–specificity (false positive rate) at the threshold applied. The fitted HSROC curve (black line) illustrates the overall trade-off between sensitivity and specificity across different CRP thresholds.

The distribution of study points shows clear threshold effects: lower thresholds (e.g. CRP ≥ 20 mg/L) are associated with higher sensitivity but reduced specificity, while higher thresholds (e.g. CRP ≥ 100 –200 mg/L) shift the balance towards greater specificity at the expense of sensitivity. The curve's shape indicates that CRP provides only moderate discriminative ability, with no threshold offering both very high sensitivity and high specificity.

This suggests that while CRP can help rule out pneumonia at lower thresholds (useful for safety in ambulatory care), it does so at the cost of many false positives. Conversely, higher thresholds may help rule in pneumonia with more certainty but risk missing cases. Clinically, this means CRP may be best applied as part of a triage or decision-support strategy rather than a stand-alone diagnostic test.

Remark:

HSROC models do not typically report AUC as standard output. You can derive it if needed, but guidance (e.g. Cochrane Handbook, Chapter 10) does not recommend reporting AUC as a primary outcome, because it can be misleading. AUC summarises accuracy across all possible thresholds, but in practice, only clinically relevant thresholds matter, and these are better illustrated directly on the HSROC curve or by presenting threshold-specific summary estimates.

6.6.4 Analysis tools for meta-analysis in DTA reviews

6.6.4.1 RevMan plots with MetaDTA or R

RevMan can generate plots for diagnostic test accuracy meta-analyses, but it does not perform the statistical analyses itself. Instead, results from analyses carried out in other software must be entered into RevMan in order to create the plots.

There are several options to conduct the meta-analyses. Here are some examples of the applications used in this manual:

- Online application MetaDTA: Diagnostic Test Accuracy Meta-Analysis v2.1.5. see [website](#).
- R packages mada (Bivariate) and diagmeta (HSROC)

The MetaDTA Shiny App is a valuable tool for DTA meta-analyses, particularly for those who prefer a user-friendly, non-programming environment. Its intuitive interface makes the analysis process easier, and it can produce a 'Parameters for RevMan' file, which is necessary for integrating the results into Cochrane reviews. However, the app is less flexible and customisable than performing the same analyses directly with R packages, which require programming skills but allow for greater control over the models and outputs.

6.6.4.2 Bivariate meta-analysis using 'mada'

The mada package implements bivariate models, simultaneously analysing both sensitivity and specificity, while accounting for the inherent correlation between these two measures.

It fits a bivariate Random-Effects Model: the Reitsma et al. (2005) model.⁷⁸ This model provides a robust framework for pooling data and generating a summary estimate of diagnostic accuracy. The central function for this is `reitsma()`. Mada is designed to work with the standard 2x2 table data from DTA studies (True Positives, False Negatives, False Positives, and True Negatives). It also includes functions to handle descriptive statistics, calculate diagnostic odds ratios, and perform meta-regression to investigate the impact of study-level covariates.

Example R-code

```
library(readxl)

datacrp <- read_excel("C:/Users/...../Pneumonia_CRP.xlsx") #enter the location of the dataset on your
computer

View(datacrp)

install.packages("mada")

library(mada)

datacrp <- as.data.frame(datacrp)

reitsma(data = datacrp, TP="TP", FN="FN", FP="FP", TN="TN", subset = NULL, method = "reml")

plotcrp <- reitsma(data = datacrp, TP="TP", FN="FN", FP="FP", TN="TN", subset = NULL, method =
"reml")

plot(plotcrp,
      ciSens = TRUE, ciSpec = TRUE, mark.optcut = FALSE,
      shading = "hatch")
```



6.6.4.3 HSROC analysis using 'diagmeta'

The diagmeta package in R was developed specifically to implement the methods introduced by Steinhauser and colleagues (2016, *Stat Med*) for diagnostic test accuracy meta-analysis with multiple thresholds: it fits a hierarchical model that jointly analyses data from studies reporting test performance at multiple thresholds; it estimates sensitivity and specificity at pre-specified or data-driven thresholds, rather than forcing a single summary point; it produces HSROC curves based on these threshold-specific estimates; and it can handle both single-threshold and multi-threshold studies, which is important for tests where many cut-offs are reported (e.g. CRP in the examples).⁷⁹

Example R-code

```
library(readxl)

datacrp <- read_excel("C:/Users/...../Pneumonia_CRP.xlsx") #enter the location of the dataset on your
computer
View(datacrp)
install.packages("diagmeta")
datacrp <- as.data.frame(datacrp)
library(diagmeta)

#different model assumptions possible:
#diagmeta(TP, FP, TN, FN, cutoff, study, data = datacrp, model = "CS")
#diagmeta(TP, FP, TN, FN, cutoff, study, data = datacrp, model = "CI")
#diagmeta(TP, FP, TN, FN, cutoff, study, data = datacrp, model = "DS")
#diagmeta(TP, FP, TN, FN, cutoff, study, data = datacrp, model = "CICS")
#diagmeta(TP, FP, TN, FN, cutoff, study, data = datacrp)
diagmeta(TP, FP, TN, FN, cutoff, study, data = datacrp, model = "CIDS")
plot1 <- diagmeta(TP, FP, TN, FN, cutoff, study, data = datacrp, model = "CS")
plot1 <- diagmeta(TP, FP, TN, FN, cutoff, study, data = datacrp, model = "CIDS")
plot.diagmeta(plot1)

plot(plot1, which = c("sroc"), col.points = "blue", pch.points = 20, cex = 1.5, main = "CRP for Pneumonia",
ciSens = FALSE,

      ciSpec = FALSE, col.ci = "blue", mark.optcut = FALSE, col.optcut = "red", mark.cutpoints = TRUE,
cex.marks = 0.8, shading = "none")
```

Model assumptions

The different model assumptions in the example R-code (CS, CI, DS, CICS, CIDS) refer to the structure of the linear mixed-effects model used to fit the transformed data.

The letters in each acronym stand for:

- **C** = Common
- **D** = Different
- **I** = Intercept
- **S** = Slope

These terms describe the assumptions about the random effects in the model, which account for the between-study heterogeneity in the data:

- CS (Common Slope)

This model assumes that the slope of the transformed distribution function is the same for both the diseased and non-diseased groups across all studies. It implies that the relationship between the biomarker values and the transformed sensitivity/specificity is consistent, but the "location" of the distributions (the intercept) can vary between studies.

- CI (Common Intercept)

This model assumes that the intercept of the transformed distribution function is the same for both the diseased and non-diseased groups across all studies. This implies that the location of the distributions is consistent, but the relationship between the biomarker and the transformed accuracy (the slope) can vary between studies.

- DS (Different Slope)

This model allows the slope of the transformed distribution function to be different for the diseased and non-diseased groups, and these slopes can vary from study to study. The intercept, however, is assumed to be common. This is a more flexible model than CS.

- CICS (Common Intercept, Common Slope)

This is a highly restrictive model. It assumes that both the intercept and the slope of the transformed distribution function are the same for the diseased and non-diseased groups across all studies. It implies very little between-study heterogeneity.

- CIDS (Common Intercept, Different Slope)

This model assumes a common intercept but allows the slopes to be different for the diseased and non-diseased groups. This is the same as the DS model. It accounts for a key aspect of heterogeneity, where the relationship between a biomarker and its accuracy can vary across different studies, even if the baseline performance (intercept) is similar.

These models are essentially different ways to account for between-study variability and the relationship between sensitivity and specificity as the cut-off value changes. By comparing the fits of these models, researchers can determine the most appropriate assumptions for their specific meta-analysis.

The CICS (Common Intercept, Common Slope) model is often the default assumption. This is because it is the simplest and most parsimonious model, requiring the least number of parameters to be estimated. However, researchers often test other more complex models like CS, CI, DS, CIDS, and DICS (Different Intercept, Common Slope) to account for varying levels of between-study heterogeneity. The choice of the best model depends on the data set being analysed and is often determined by statistical tests or model fit criteria, such as the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC).

Following YouTube video provides a practical guide on how to use the package and its functionalities: see [website](#).

6.7 Rating the certainty of evidence (Update 2025)

The GRADE working group described in several recent papers how to address the certainty of evidence from test DTA studies, and how to present the results of an assessment.^{67, 68, 74, 81}

As in GRADE for interventions, the certainty ratings reflect the extent of our confidence that the estimates of an effect are adequate to support a particular decision or recommendation.³⁵ GRADE categorises the certainty of the evidence in four levels: high, moderate, low, and very low. that are applied to a body of evidence, and not to individual studies.

The type of estimate for which certainty is assessed varies depending on the application of GRADE; in diagnostic questions, it may reflect the effect of a diagnostic strategy or, more commonly, the diagnostic accuracy measures of a test.



In the context of a DTA systematic review, certainty reflects our confidence that the accuracy estimates are correct. Factors that decrease the certainty of evidence for studies of diagnostic accuracy involve the study design, the risk of bias (limitations in study design and execution), the presence of indirectness (for the patient population, the diagnostic test, the comparison test and indirect comparisons of tests), an important Inconsistency in study results, imprecise evidence, and high probability of publication bias.

DTA cross-sectional or cohort studies in patients with diagnostic uncertainty and direct comparison of test results with an appropriate reference standard (current best possible alternative test strategy) are considered high certainty. Thus, GRADE considers DTA studies with an appropriate cross-sectional design or cohort studies, of initial high certainty of evidence (in contrast to the same study design in GRADE for interventions)(see Figure 4).

Upgrading of the certainty may be indicated for dose effect, large effects residual plausible bias and confounding. These factors are further discussed in the next paragraphs.

6.7.1 Study limitations, risk of bias

For grading, the risk of bias part of QUADAS-2 (see section 4.4.1) is used for assessing study limitations across the body of evidence. As for interventions, it is advisable to report risk of bias assessment separately for different outcomes, or report transparently that the risk of bias is considered similar for all outcomes.

Study limitations in DTA studies are assessed separately for patients with the target condition (sensitivity estimates; TP and FN) and patients without the target condition (specificity estimates; TN and FP).

The criteria that should be taken into account are:

- Representativeness of the population that was intended to be sampled
- Independent comparison with the best alternative test strategy
- All enrolled patients should receive the new test and the best alternative test strategy
- Diagnostic uncertainty should be given
- Is the reference standard likely to correctly classify the target condition?

Moving from risk of bias of individual studies to a judgement about rating down for study limitations across a body of DTA evidence mainly relies on judgement and is challenging. Often, studies of test accuracy suffer very serious methodological limitations that warrant downgrading the certainty of evidence by one or two levels; from high, through moderate, to low.⁶³

Examples

Patient selection: lack of including a consecutive sample (or a random sample) of patients or the use of a case-control design for assessing the DTA of a test; patient population included should represent the intended inclusion PICO criteria;

Index test: interpreting the index test with knowledge of the results of the reference standard (independent comparison) or not having used a pre-specified threshold for positivity of the index test; all participants should have the same index and reference standard;

Reference standard (best alternative test strategy): having used a reference standard that may not classify patients correctly as having or not having the target condition or interpretation of the reference standard with knowledge of the results of the index test;

Flow and timing: inappropriate interval between conducting the index test and the reference standard, applying the reference standard to a subset of patients only (partial verification), not having applied the

same reference standard to all patients (differential verification) or not having included all patients in the analysis;

In addition, due to incomplete reporting of DTA studies, items might have been scored 'Unclear', hampering a sound judgement of the risk of bias. So, a judgement must be made about whether or not bias is expected for the estimated summary estimates of both sensitivity and specificity due to study limitations across studies (and not individual studies).

6.7.2 Indirectness

In the DTA domain, indirectness can be defined on several levels: indirectness on a population, test intervention, test comparison, and outcome level.

The indirectness on a population level – the patient selection

The applicability domain of QUADAS-2 patient selection evaluates if there is concern that the included participants do not match the review question (see section 4.4.2.6). Indirect evidence results from differences in the disease spectrum and the setting, or when prior testing or referral paths influenced the population tested. Prevalence and pre-test probabilities in the studies can be consulted to guide the evaluation of indirectness. Sensitivity analyses can help define the sources of variability in the available evidence.

The indirectness on the test intervention – index test

The applicability domain of QUADAS-2 index test evaluates if there concerns that the index test, its conduct, or interpretation differ from the review question (see section 4.4.3.6). Indirect evidence results from differences in the standards used, the cut-off settings, the specimen transfer (influence of transport conditions), and general technical variability and operator reliability.

The indirectness on the test comparison – reference standard

The applicability domain of QUADAS-2 evaluates if there are concerns that the reference standard, its conduct, or interpretation differ from the review question (see section 4.4.4.6). DTA studies with more than one comparator, e.g. with a test of non-interest indirectness will result, or when an indirect comparison is made of 2 tests used in separate studies (with the same reference standard) and not directly compared in the same studies.

The indirectness on the test intervention – outcomes

GRADE regards DTA studies mostly as low certainty evidence because of the absence of direct evidence about impact on patient-important outcomes.⁶⁹ A GDG must make deductions from the DTA studies about the balance between the presumed influences on patient-important outcomes of any differences in true and false positives and true and false negatives taking test complications and costs into account. Therefore, accuracy studies typically provide low certainty evidence for making recommendations due to indirectness of the outcomes, similar to surrogate outcomes for treatments.

A summary table for indirectness judgements across the body of evidence by TP etc. is proposed by GRADE and can be used to come to a final judgement (see Tabel 4 of GRADE guidelines 21 part 1).⁶⁸

6.7.3 Inconsistency, heterogeneity

6.7.3.1 Definitions of heterogeneity in DTA

In DTA reviews, the terms clinical, methodological, and statistical heterogeneity are often mentioned together, but they describe different aspects of variability (Table 17).

- *Clinical heterogeneity* refers to differences in the patients, settings, or test use across studies. It is important to assess those differences, because they affect how generalisable the results are. Examples:



- Study populations: adults versus children, primary care versus emergency department.
- Disease spectrum: mild versus severe cases, different prevalence.
- Differences in test conduct: finger-prick versus venous blood for CRP.
- *Methodological heterogeneity* refers to differences in study design, conduct, or quality. It is important to assess those differences, because poor methods may bias sensitivity/specificity estimates. Examples:
 - Reference standard used (e.g. clinical diagnosis versus microbiological culture).
 - Blinding versus no blinding of index test interpreters.
 - Case–control versus cohort design.
 - Selective inclusion/exclusion of patients.
- *Statistical heterogeneity* refers to the observed variability in results (sensitivities, specificities, likelihood ratios) beyond what would be expected by chance. It is important to assess those differences, because it affects the validity of pooled estimates. Examples:
 - Wide spread of sensitivities/specificities at the same threshold.
 - Inconsistency not explained by clinical or methodological factors.
 - Quantified in meta-analysis with measures like variance components, prediction intervals, or visual scatter in ROC space.

Table 17 – Comparison of sources of heterogeneity in DTA reviews

Type of heterogeneity	What it means	Typical sources	Example in DTA reviews
<i>Clinical</i>	Differences in patient characteristics, setting, or test application	Population, setting (GP vs ED), spectrum of disease, index test used	CRP threshold studies in ED versus primary care give different accuracy
<i>Methodological</i>	Differences in study design, conduct, or quality	Reference standard, blinding, patient selection, study design	Using chest X-ray vs microbiological culture as the reference standard
<i>Statistical</i>	Observed variability in results not due to chance	Random error + residual differences not explained by clinical/methodological factors	Wide variation in sensitivity across studies even with same CRP cut-off

When we see statistical heterogeneity (e.g. spread in Forest or ROC plots), we have to look backwards at possible clinical and methodological explanations. Clinical heterogeneity is interconnected with methodological heterogeneity. Both lead to statistical heterogeneity (observed results).

6.7.3.2 When is heterogeneity “too large” to proceed with a meta-analysis?

There are no absolute statistical thresholds for deciding that heterogeneity is excessive in DTA meta-analysis (unlike I^2 in intervention reviews, which itself is imperfect). Instead, heterogeneity is too large to pool if:

- studies clearly address different review questions (different populations, thresholds, or reference standards). Pooling would be misleading, as it produces a summary estimate that does not represent any real-world context.

- there are too few studies: hierarchical models (bivariate or HSROC) generally need ≥ 4 studies, and preferably ≥ 10 , to reliably estimate between-study variance. With fewer, estimates may be unstable, especially if heterogeneity is obvious.
- we observe extremely inconsistent accuracy estimates: e.g. some studies showing sensitivities near 100% and others near 20%, without an identifiable explanation (thresholds, populations, etc.). In such cases, present results descriptively (forest/ROC plots) rather than attempting pooled estimates.

Practical guidance (from the Cochrane Handbook for DTA Reviews):

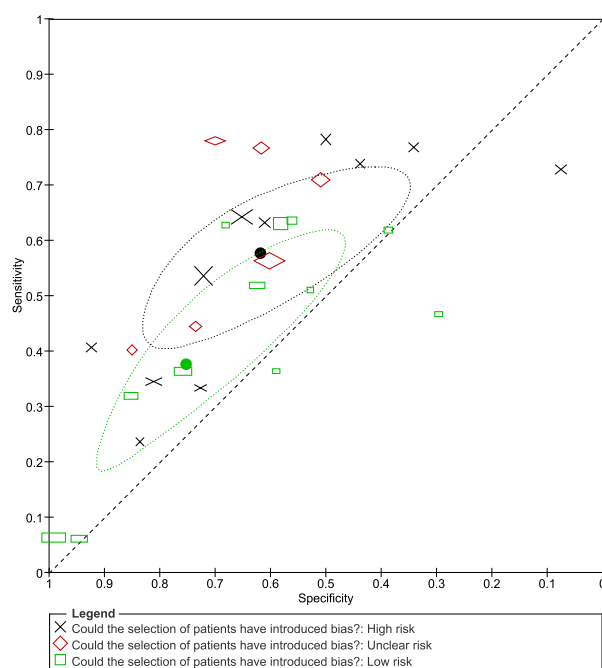
- Do meta-analysis when: studies are sufficiently similar in clinical question, use comparable thresholds/reference standards, and provide ≥ 4 datasets with consistent enough results to justify pooling.
- Do not meta-analyse when: heterogeneity is extreme, unexplained, or when studies are too sparse. In these cases, descriptive synthesis (forest plots, ROC plots) is more appropriate.

6.7.3.3 Exploring heterogeneity by the use of covariates

We strongly recommend that the exploration of heterogeneity be pre-specified in the protocol of a DTA review, including an explicit statement of which covariates will be examined. Although it may be tempting, it is not good practice to investigate every possible covariate encountered during the review in the meta-regression.

Exploring heterogeneity in a DTA review using covariates involves extending the bivariate random-effects model to a bivariate meta-regression. This allows assessment of whether study characteristics (setting, test definitions, thresholds, populations, or risk of bias) explain variation in sensitivity and specificity. Results are typically presented as stratified estimates in tables or subgroup summary points in ROC plots, helping reviewers understand how and why diagnostic accuracy varies. Plots are easier to interpret than estimates in a table.

Figure 18 – Diagnostic value of fever for COVID-19: bivariate meta-regression by risk of bias of participant selection (QUADAS-2)



Source: Struyf T et al. ⁸⁰



Figure 18 shows the pooled estimates for fever, based on the risk of bias for participant selection. The estimates for a low risk of bias are represented by the green rectangles and the green summary point (12 studies). Estimates for a high risk of bias are represented by the black crosses and the black summary point (11 studies). No summary point was estimated for studies with an unclear risk of bias, because this category is difficult to interpret and contained too few studies (n=6). The two ellipses are the 95% confidence regions around each subgroup summary point.

The covariate (risk of bias in patient selection) clearly explains some variation:

- Low risk studies suggest that fever is not very sensitive for COVID-19 diagnosis but is relatively specific when present (rule-in, not rule-out).
- Higher/unclear risk studies report better sensitivity but worse specificity, which is likely due to spectrum bias. For example, selective recruitment of highly symptomatic patients in hospitals may artificially inflate sensitivity.

This pattern is consistent with known biases: studies that do not carefully avoid selection bias often overestimate sensitivity.

6.7.3.4 Sensitivity analysis

In bivariate meta-analysis, sensitivity analyses are essential for assessing the robustness of results. Common approaches include excluding outliers, restricting analyses to studies judged at low risk of bias, and focusing on subgroups that use consistent thresholds or clinical settings. Leave-one-out analyses can be used to evaluate the stability of summary estimates by checking whether findings change when individual studies are removed.⁸²

In addition, testing alternative model assumptions (e.g. different correlation structures or variance estimations) helps ensure that conclusions are not unduly dependent on specific analytical choices. Together, these checks provide reassurance that pooled estimates of sensitivity and specificity are reliable and not driven by a small number of atypical or biased studies.

6.7.3.5 Downgrading for inconsistency

Analogous to the therapeutic domain the certainty of the evidence should be graded down by one or two levels (depending on the magnitude of the inconsistency in the results) when heterogeneity exists for which investigators were not able to identify a plausible explanation. In meta-analysis of DTA studies, however, heterogeneity (or in GRADE terminology 'inconsistency') is the rule rather than the exception. The estimates for sensitivity and specificity should be rated separately.

Because in DTA meta-analyses mainly random effects models are used, inconsistency (heterogeneity between studies), but also by imprecision due to small study sizes (like in fixed effect models) might be present. One should not automatically grade down twice (see explanation in 6.7.4).

6.7.4 Imprecision

The assessment of the confidence intervals of absolute test accuracy results (true and false, positive, and negative results) remains the cornerstone for evaluating imprecision. For rating the certainty of evidence the target or the decision thresholds should be known. These should have been defined at the start of the project (see section 6.4.2). The judgement is based on the 95% confidence intervals around the two categories of test results that have the most important consequences for patients: the number of FPs and the number of FNs. These 95% confidence intervals are derived from the 95% confidence intervals of the summary sensitivity and specificity and are presented in a SoF Table, which includes the summary 2*2 Table in a hypotheticalal population of 1000 tested.

Consider rating down certainty of evidence by one or more levels, when a CI crosses one or more predefined threshold(s). When no threshold is crossed, the effect is large and the number of events are small, fragility might be present (see section 5.5.5.3 for the explanation of fragility and optimal

information size and review information size). To evaluate this, the sample size for an adequately powered test accuracy (optimal information size (OIS)) review should be calculated.⁷⁴ The description and [link](#) for performing this calculation is in the appendix by Mustafa et al 2024.⁷⁴

For the calculation the following is required

- The prevalence of the target condition;
- The sensitivity/specificity from the systematic review;
- The acceptable margin of error defined
 - As an arbitrary difference (e.g., 0.10 difference in sensitivity), or
 - back calculated from FPs and FNs proportions that have practical consequences to stakeholders. The margin of error can be derived from the difference between these thresholds so that calculation informs whether the analysis is sufficiently powered to discriminate between such differences.

If the combined sample size of the included studies in the review is smaller than the required OIS/RIS, one should consider rating down by one or more levels for imprecision.

The relationship between inconsistency and imprecision

GRADE guidance 37 provides the steps for grading inconsistency and imprecision.⁷⁴ In random effects meta-analysis, the CI of the summary estimate is derived from both within and between study heterogeneity. Large between-study heterogeneity can result in a wide CI of the summary estimate. This may lead to rating down certainty twice, once for inconsistency and once for imprecision. This double rating down may be inappropriate. An example of different scenarios is presented in the guidance 37.⁷⁴

Steps from GRADE 37:

1: Check CI for imprecision: summary estimate crossing a threshold?

- Important imprecision: YES

2: Check for inconsistency (see section 6.7.3 for explanation; avoid I^2 which may lead to misleading conclusions)

- Important inconsistency: YES

3: Evaluate the OIS

- If the available sample size across studies is lower than OIS, then it would be appropriate to lower certainty for both domains
- If the sample size exceeds OIS, this may be a situation when we opt to not rate down for both domains and only rate down for inconsistency; provide an explanation that, though the presence of some imprecision was recognised, imprecision was attributed to inconsistency

Lastly, when CIs of summary estimates cross multiple judgement thresholds, imprecision is likely severe, and it would be appropriate to rate down for both domains.

If the sample size exceeds OIS and CIs of FP/FN proportions do not overlap test accuracy thresholds, then there is no need to lower certainty for imprecision.

6.7.1 Publication bias in DTA reviews

Time to publication of diagnostic accuracy studies with promising results about the performance of tests are published more rapidly compared to those reporting lower estimates.⁸³

Assessment of publication bias in DTA reviews is more complex than in intervention reviews. Standard funnel plots are not appropriate, as DTA outcomes are two-dimensional (sensitivity and specificity)



and often highly heterogeneous. An adapted method, Deeks' funnel plot, plots the log diagnostic odds ratio against the inverse square root of the effective sample size, with a regression test for asymmetry.^{84, 85} The Cochrane Handbook for DTA reviews recommends this approach when there are sufficient studies (generally ≥ 10).

However, funnel plots are seldom used in practice: most DTA reviews include too few studies, threshold effects and heterogeneity can mimic small-study bias, and interpretation is uncertain. Thus, visual or statistical evidence of asymmetry should be reported cautiously, and absence of asymmetry does not exclude publication bias.

RevMan does not provide funnel plots for DTA reviews. Where ≥ 10 studies are available, Deeks' funnel plot can be produced in external software (e.g. Stata or R), but results should always be interpreted with caution.

Further, Cochrane encourages to explore the mechanisms underpinning reporting bias in the studied clinical area and advises review authors to explain why any detected publication bias may be problematic.

At KCE, typically, language bias should be reduced by searching broader than in English e.g. French, German, Spanish, Dutch. Secondly, the KCE information specialist should be consulted to avoid flawed search strategies.

In summary, publication bias is difficult to assess for DTA studies and grading down should only be done if the evidence is convincing, e.g. when published evidence is limited to few small studies, in particular, if they support a presumed hypothesis and were funded by a body with a vested interest in a particular diagnostic method.⁸⁶

7 RECOMMENDATIONS

7.1 Evidence to decision framework (Update 2025)

The transitioning from evidence to recommendations using a structured, transparent process is described by GRADE (EtD framework) and applied at KCE (see Core GRADE paper 7).^{27, 65, 87}

Guideline panels usually provide a single grade of certainty of evidence for every recommendation, but the strength of a recommendation usually depends on evidence regarding not just one, but a number of patient-important outcomes and on the certainty of evidence for each of these outcomes.

Formulating a recommendation (even if the level of evidence is low) should always be the aim and not formulating a recommendation an exception. Formulating an ungraded recommendation, named a 'Good practice statement' is an option (see 7.3).

The process at KCE is as follows:

- The KCE team presents the summary of findings tables with the GRADE certainty of evidence for each critical and important outcome to the GDG. The minimal important difference for each outcome and the range of effects (i.e. a trivial, small, moderate or large effect) that is bounded by two thresholds is taking into account for decision making.^{33, 34}
- The following aspects on the intervention or strategy under study are summarised by the KCE team and presented; the GDG and/or patient representatives are consulted for their input on the following aspects:
 - the benefits and harms for the patient
 - the burdens for the patient
 - the patients' values and preferences
 - the estimated cost or resource utilisation (for the patient and society)
 - feasibility (for the patient and society)
 - acceptability (for the patient and society)
 - equity (for the patient and society)

Ideally, the judgements for these criteria are supported by research evidence (systematic reviews). If this is considered too resource intensive, this should be stated.

- The direction of the recommendation - for or against - is finalised by the KCE team and GDG.
- The strength of the recommendation - strong or conditional- based on the certainty of the supporting evidence and the aspects listed above is finalised by the KCE team and GDG. The strength of recommendation according to the 4 categories of the GRADE system is presented in Table 18. The influence of aspects on strength is presented in Table 19. The interpretation of strong and conditional recommendations will be different by clinicians, patients or policy makers. This difference is explained in Table 20.

More in depth information on patients' values and preferences is available in the GRADE guidelines 19 and 20.^{88, 89} For cost-effectiveness considerations we refer to the GRADE guidance 23.⁹⁰ The equity considerations are explored by the GRADE group in Dewidar et al 2025 and Pottie et al 2017 with an example presented by Dewidar et al 2024.⁹¹⁻⁹³

**Table 18 – Strength of recommendation according to the 4 categories of the GRADE system**

Grade	Definition
Strong for	Recommendation for The desirable effects of an intervention clearly outweigh the undesirable effects (<i>the intervention is to be put into practice</i>)
Strong against	Recommendation against The undesirable effects of an intervention clearly outweigh the desirable effects (<i>the intervention is not to be put into practice</i>)
Conditional for	Recommendation for The desirable effects of an intervention probably outweigh the undesirable effects (<i>the intervention probably is to be put into practice</i>)
Conditional against	Recommendation against The undesirable effects of an intervention probably outweigh the desirable effects (<i>the intervention probably is not to be put into practice</i>)

Source: 27, 87

Table 19 – Factors that influence the strength of a recommendation

Factor	Comment
Balance between clinical benefit and harms	The larger the difference between the desirable and undesirable effects, the higher the likelihood that a strong recommendation is warranted. The narrower the gradient, the higher the likelihood that a conditional recommendation is warranted
Certainty of evidence	The higher the certainty of evidence, the higher the likelihood that a strong recommendation is warranted
Values and preferences	The more values and preferences vary, or the greater the uncertainty in values and preferences, the higher the likelihood that a conditional recommendation is warranted
Costs (resource allocation)	The higher the costs of an intervention, i.e. the greater the resources consumed, the lower the likelihood that a strong recommendation is warranted

Sources: 27, 87, 94

Table 20 – Interpretation of strong and conditional recommendations by audience

Implications	Strong recommendation	Conditional recommendation
For patients	Most individuals in this situation would want the recommended course of action, and only a small proportion would not. Formal decision aids are not likely to be needed to help individuals make decisions consistent with their values and preferences.	The majority of individuals in this situation would want the suggested course of action, but many would not.
For clinicians	Most individuals should receive the intervention. Adherence to this recommendation according to the	Recognise that different choices will be appropriate for individual patients and that you must help each patient arrive at a management decision consistent with his or

Implications	Strong recommendation	Conditional recommendation
	guideline could be used as a quality criterion or performance indicator.	her values and preferences. Decision aids may be useful helping individuals making decisions consistent with their values and preferences.
For policy makers	The recommendation can be adopted as policy in most situations.	Policy-making will require substantial debate and involvement of various stakeholders.

Source: ⁹⁴

7.2 Wording the strength of a recommendation (Update 2025)

This topic is particular for KCE as three languages are used in our reporting. Recommendations are first drafted in English in the scientific report. In the next stage, and after validation of the scientific report, a French and a Dutch synthesis with the recommendations are published. This translation is a painstaking work as nuances are often difficult to capture correctly.

For the most recent GCP-report on melanoma, researchers decided on the use of ‘conditional’ over ‘weak’. The reasoning being that the concept of “weak” can be interpreted as “unimportant” but this is far from the point.^{94, 95}

We present here the wording in the three languages that KCE leans towards for the titles (Table 21) and for the wording to be used in the recommendation itself (Table 22). Although literature points to using “direct” and “actionable” styles in formulating recommendations, this might be experienced as paternalistic in French or Dutch. The preferred statements at KCE in Table 22 are indicated in blue / shaded.

Table 21 – Preference KCE for ‘strong and conditional’ in French and Dutch

Strength of the recommendation	Force de la recommandation	Kracht van de aanbeveling
Preference KCE		
Strong	Forte	Sterk
Conditional	Conditionnelle	Voorwaardelijk
Second choice		
Weak	Faible	Zwak
Discretionary	A envisager au cas par cas	Te overwegen afhankelijk van de context
Qualified	Recommandation avec réserves	Aanbeveling met kanttekeningen

Table 22 – Preference KCE for wording of recommendations in French and Dutch

English	French	Dutch
Strong recommendation in favour		
<i>Clinicians should apply X</i>	X devrait être réalisée	X wordt toegepast / uitgevoerd / voorgesteld / voorgeschreven
<i>We recommend X</i>	Nous recommandons X	We bevelen aan X
	Les cliniciens devraient	Artsen (zouden) moeten



Weak recommendation in favour		
<i>Clinicians may apply X</i>	X peut être réalisé / prescrit / assuré / proposé	X kan worden toegepast / uitgevoerd / voorgesteld / voorgeschreven
<i>We suggest</i>	Nous suggérons	We suggereren
<i>Clinicians might</i>	Les cliniciens pourraient	Artsen zouden kunnen
<i>We conditionally recommend</i>	Nous recommandons sous condition	We bevelen aan te overwegen
Strong recommendation against		
<i>Clinicians should not offer X</i>	X ne devrait pas être proposé / utilisé	X wordt niet aanbevolen
<i>We recommend not ..</i>	Nous recommandons de ne pas ...	We bevelen aan om... niet
Weak recommendation against		
<i>We suggest ...not</i>	Nous suggérons de ne pas ...	We suggereren om niet ...
<i>Clinicians might not ...</i>	Les cliniciens pourraient ne pas ...	Artsen zouden mogelijk niet ...
<i>We conditionally recommend ...not</i>	Nous recommandons sous condition de ne pas ...	We bevelen aan niet te overwegen ...

X= the intervention; example in blue / shaded is preferred by KCE. The word clinician can stand for any healthcare worker.

7.3 Formulating Good practice statements (Update 2025)

The GRADE Working Group developed ungraded best or good practice statements (GPS) for those situations when a guideline panel might judge the certainty high that the desirable effects of an intervention clearly outweighs its undesirable effects but the body of supportive evidence is indirect (as was the case during the COVID pandemic when the GPS were first developed).²⁹ In such cases, the application of the GRADE approach for grading the strength of recommendations is inappropriate.²⁹

The following criteria should be followed when considering a GPS:

- the statement should be clear and actionable;
- the message should be necessary for actual healthcare practice (i.e., without the guidance, clinicians would fail to make the appropriate decision);
- after considering all relevant outcomes and potential downstream consequences, implementing the GPS would result in large net positive consequences. This should be assessed (informally) against as many of the regular EtD criteria as possible;
- collection and summary of the evidence would be a poor use of a guideline panel's time and resources; and
- a well-documented, clear and explicit rationale connecting the indirect evidence should be constructed.

For an example of applying GPS, see the paper by Dewidar et al.²⁹

8 REVIEW AND VALIDATION

8.1 Stakeholder review

At the end of the project, the results are presented to the stakeholders in a meeting or survey. A draft of the final recommendations is reviewed for clarity of the wording, feasibility and acceptability for patients, socio-cultural acceptability, feasibility for the health system, and health equity. The uptake of recommendations is also discussed.

8.2 Validation by external experts

An independent external review of the scientific report is performed by three experts (i.e. clinical expert(s) and a methodological expert(s)) who were not involved in the GDG. This validation aims to improve the scientific value of the report.

The validation includes a complete and thorough assessment of the guideline and appendices using the AGREE II instrument. This is presented by the KCE team as part of the guideline.

9 DISSEMINATION

A range of different methods to raise awareness of the guideline exist. These include standard approaches which are applied by KCE, such as:

- issuing a press release at the moment of publication on the KCE website;
- addressing questions from the press;
- notifying experts, stakeholders and patient organisations of the publication;
- publicising the guideline through the KCE's alerts;
- using social media channels;
- and publicising the guideline's main points of interest in a journal.

For the applicability of the guideline, we refer to the section on evidence to recommendation which discusses facilitators and barriers, such as resources, cost or patient's opinions on the topic under study.



10 MANUAL KCE - VALIDATED ASSESSMENT TOOLS

10.1 AGREE II – guidelines

The AGREE II evaluation score is used to critically appraise guidelines (Table 23). Each of the AGREE II items are rated on a 7-point scale (1–strongly disagree to 7–strongly agree). The User's Manual provides guidance on how to rate each item using the rating scale. Website [AGREE](#) and the [AGREE checklist](#).

Table 23 – AGREE II instrument

Domain 1. Scope and Purpose
1. The overall objective(s) of the guideline is (are) specifically described.
2. The health question(s) covered by the guideline is (are) specifically described.
3. The population (patients, public, etc.) to whom the guideline is meant to apply is specifically described.
Domain 2. Stakeholder Involvement
4. The guideline development group includes individuals from all the relevant professional groups.
5. The views and preferences of the target population (patients, public, etc.) have been sought.
6. The target users of the guideline are clearly defined.
Domain 3. Rigour of Development
7. Systematic methods were used to search for evidence.
8. The criteria for selecting the evidence are clearly described.
9. The strengths and limitations of the body of evidence are clearly described.
10. The methods for formulating the recommendations are clearly described.
11. The health benefits, side effects, and risks have been considered in formulating the recommendations.
12. There is an explicit link between the recommendations and the supporting evidence.
13. The guideline has been externally reviewed by experts prior to its publication.
14. A procedure for updating the guideline is provided.
Domain 4. Clarity of Presentation
15. The recommendations are specific and unambiguous.
16. The different options for management of the condition or health issue are clearly presented.
17. Key recommendations are easily identifiable.
Domain 5. Applicability
18. The guideline describes facilitators and barriers to its application.
19. The guideline provides advice and/or tools on how the recommendations can be put into practice.
20. The potential resource implications of applying the recommendations have been considered.
21. The guideline presents monitoring and/ or auditing criteria.
Domain 6. Editorial Independence
22. The views of the funding body have not influenced the content of the guideline.
23. Competing interests of guideline development group members have been recorded and addressed.

10.2 AMSTAR 2 – systematic reviews

The AMSTAR-2 was published in 2017 by Shea et al.¹²

Table 24 – Evaluation criteria AMSTAR-2 for a systematic review

Did the research questions and inclusion criteria for the review include the components of PICO?		
For Yes: ☐ Population ☐ Intervention ☐ Comparator group ☐ Outcome	Optional (recommended) ☒ Timeframe for follow-up	☐ Yes ☐ NO
Did the report of the review contain an explicit statement that the review methods were established prior to the conduct of the review and did the report justify any significant deviations from the protocol?		
For Partial Yes: The authors state that they had a written protocol or guide that included ALL the following: ☐ review question(s) ☐ a search strategy ☐ inclusion/exclusion criteria ☐ a risk of bias assessment	For Yes: As for partial yes, plus the protocol should be registered and should also have specified: ☐ a meta-analysis/synthesis plan, if appropriate, and ☐ a plan for investigating causes of heterogeneity ☐ justification for any deviations from the protocol	☐ Yes ☐ Partial Yes ☐ No
Did the review authors explain their selection of the study designs for inclusion in the review?		
For Yes, the review should satisfy ONE of the following: ☐ Explanation for including only RCTs ☐ OR Explanation for including only NRSI ☐ OR Explanation for including both RCTs and NRSI		☐ Yes ☐ No
Did the review authors use a comprehensive literature search strategy?		
For Partial Yes (all the following): ☐ searched at least 2 databases (relevant to research question) ☐ provided key word and/or search strategy ☐ justified publication restrictions (e.g. language)	For Yes, should also have (all the following): ☐ searched the reference lists / bibliographies of included studies ☐ searched trial/study registries ☐ included/consulted content experts in the field ☐ where relevant, searched for grey literature	☐ Yes ☐ Partial Yes ☐ No



	☐ conducted search within 24 months of completion of the review	
Did the review authors perform study selection in duplicate?		
For Yes, either ONE of the following: ☐ at least two reviewers independently agreed on selection of eligible studies and achieved consensus on which studies to include ☐ OR two reviewers selected a sample of eligible studies and achieved good agreement (at least 80 percent), with the remainder selected by one reviewer.		☐ Yes ☐ No
Did the review authors perform data extraction in duplicate?		
For Yes, either ONE of the following: ☐ at least two reviewers achieved consensus on which data to extract from included studies ☐ OR two reviewers extracted data from a sample of eligible studies and achieved good agreement (at least 80 percent), with the remainder extracted by one reviewer.		☐ Yes ☐ No
Did the review authors provide a list of excluded studies and justify the exclusions?		
For Partial Yes ☐ provided a list of all potentially relevant studies that were read in full-text form but excluded from the review	For Yes, must also have: ☐ Justified the exclusion from the review of each potentially relevant study	☐ Yes ☐ Partial Yes ☐ No
Did the review authors describe the included studies in adequate detail?		
For Partial Yes (ALL the following): ☐ described populations ☐ described interventions ☐ described comparators ☐ described outcomes ☐ described research design	For Yes, should also have ALL the following: ☐ described population in detail ☐ described intervention in detail (including doses where relevant) ☐ described comparator in detail (including doses where relevant) ☐ described study's setting ☐ timeframe for follow-up	☐ Yes ☐ Partial Yes ☐ No
Did the review authors use a satisfactory technique for assessing the risk of bias (RoB) in individual studies that were included in the review?		
RCTs For Partial Yes, must have assessed RoB from ☐ unconcealed allocation, and	For Yes, must also have assessed RoB from: ☐ allocation sequence that was not truly random, and ☐ selection of the reported result from among multiple measurements or analyses of a specified outcome	☐ Yes ☐ Partial Yes ☐ No ☐ Includes only NRSI

☐ lack of blinding of patients and assessors when assessing outcomes (unnecessary for objective outcomes such as all- cause mortality) NRSI For Partial Yes, must have assessed RoB: ☐ from confounding, and ☐ from selection bias	For Yes, must also have assessed RoB: ☐ methods used to ascertain exposures and outcomes, and ☐ selection of the reported result from among multiple measurements or analyses of a specified outcome	☐ Yes ☐ Partial Yes ☐ No ☐ Includes only RCTs
Did the review authors report on the sources of funding for the studies included in the review?		
For Yes ☐ Must have reported on the sources of funding for individual studies included in the review. Note: Reporting that the reviewers looked for this information. No but it was not reported by study authors also qualifies		☐ Yes ☐ No
If meta-analysis was performed did the review authors use appropriate methods for statistical combination of results?		
RCTs For Yes: ☐ The authors justified combining the data in a meta-analysis ☐ AND they used an appropriate weighted technique to combine study results and adjusted for heterogeneity if present. ☐ AND investigated the causes of any heterogeneity For NRSI For Yes: ☐ The authors justified combining the data in a meta-analysis ☐ AND they used an appropriate weighted technique to combine study results, adjusting for heterogeneity if present ☐ AND they statistically combined effect estimates from NRSI that were adjusted for confounding, rather than combining raw data, or justified combining raw data when adjusted effect estimates were not available ☐ AND they reported separate summary estimates for RCTs and NRSI separately when both were included in the review		☐ Yes ☐ No ☐ No meta-analysis conducted ☐ Yes ☐ No ☐ No meta-analysis conducted
If meta-analysis was performed, did the review authors assess the potential impact of RoB in individual studies on the results of the meta-analysis or other evidence synthesis?		
For Yes: ☐ included only low risk of bias RCTs		☐ Yes

☐ OR, if the pooled estimate was based on RCTs and/or NRSI at variable RoB, the authors performed analyses to investigate possible impact of RoB on summary estimates of effect.	☐ No ☐ No meta-analysis conducted
Did the review authors account for RoB in individual studies when interpreting/ discussing the results of the review?	
For Yes: ☐ included only low risk of bias RCTs ☐ OR, if RCTs with moderate or high RoB, or NRSI were included the review provided a discussion of the likely impact of RoB on the results	☐ Yes ☐ No
Did the review authors provide a satisfactory explanation for, and discussion of, any heterogeneity observed in the results of the review?	
For Yes: ☐ There was no significant heterogeneity in the results ☐ OR if heterogeneity was present the authors performed an investigation of sources of any heterogeneity in the results and discussed the impact of this on the results of the review	☐ Yes ☐ No
If they performed quantitative synthesis did the review authors carry out an adequate investigation of publication bias (small study bias) and discuss its likely impact on the results of the review?	
For Yes: ☐ performed graphical or statistical tests for publication bias and discussed the likelihood and magnitude of impact of publication bias	☐ Yes ☐ No ☐ No meta-analysis conducted
Did the review authors report any potential sources of conflict of interest, including any funding they received for conducting the review?	
For Yes: ☐ The authors reported no competing interests OR ☐ The authors described their funding sources and how they managed potential conflicts of interest	☐ Yes ☐ No

Copied from http://amstar.ca/Amstar_Checklist.php



10.3 ROBIS – systematic reviews

10.3.1 Phase 1: Assessing relevance (Optional)

ROBIS is designed to assess the risk of bias in reviews with questions relating to interventions, aetiology, diagnosis and prognosis. State your overview/guideline question (target question) and the question being addressed in the review being assessed:

INTERVENTION REVIEWS:

Category	Target question (e.g. overview or guideline)	Review being assessed
<i>Patients / population(s):</i>		
<i>Intervention(s):</i>		
<i>Comparator(s):</i>		
<i>Outcome(s):</i>		

FOR AETIOLOGY REVIEWS

Category	Target question (e.g. overview or guideline)	Review being assessed
<i>Patients / population(s):</i>		
<i>Exposure(s) and comparator(s):</i>		
<i>Outcome(s):</i>		

FOR DTA REVIEWS

Category	Target question (e.g. overview or guideline)	Review being assessed
<i>Patient(s):</i>		
<i>Index test(s):</i>		
<i>Reference standard:</i>		
<i>Target condition:</i>		

FOR PROGNOSTIC REVIEWS:

Category	Target question (e.g. overview or guideline)	Review being assessed
<i>Patients:</i>		
<i>Outcome to be predicted:</i>		
<i>Intended use of model:</i>		
<i>Intended moment in time:</i>		

Does the question addressed by the review match the target question?

YES / NO / UNCLEAR



10.3.2 Phase 2: Identifying concerns with the review process

DOMAIN 1: STUDY ELIGIBILITY CRITERIA

Describe the study eligibility criteria, any restrictions on eligibility and whether there was evidence that objectives and eligibility criteria were pre-specified

- | | |
|---|--------------|
| 1. Did the review adhere to pre-defined objectives and eligibility criteria? | Y/PY/PN/N/NI |
| 2. Were the eligibility criteria appropriate for the review question? | Y/PY/PN/N/NI |
| 3. Were eligibility criteria unambiguous? | Y/PY/PN/N/NI |
| 4. Were any restrictions in eligibility criteria based on study characteristics appropriate (e.g. date, sample size, study quality, outcomes measured)? | Y/PY/PN/N/NI |
| 5. Were any restrictions in eligibility criteria based on sources of information appropriate (e.g. publication status or format, language, availability of data)? | Y/PY/PN/N/NI |

Concerns regarding specification of study eligibility criteria

Rationale for concern:

LOW/HIGH/UNCLEAR

Y=YES, PY=PROBABLY YES, PN=PROBABLY NO, N=NO, NI=NO

DOMAIN 2: IDENTIFICATION AND SELECTION OF STUDIES

Describe methods of study identification and selection (e.g. number of reviewers involved):

- | | |
|---|--------------|
| 1. Did the search include an appropriate range of databases/electronic sources for published and unpublished reports? | Y/PY/PN/N/NI |
| 2. Were methods additional to database searching used to identify relevant reports? | Y/PY/PN/N/NI |
| 3. Were the terms and structure of the search strategy likely to retrieve as many eligible studies as possible? | Y/PY/PN/N/NI |
| 4. Were restrictions based on date, publication format, or language appropriate? | Y/PY/PN/N/NI |

Concerns regarding methods used to identify and/or select studies

Rationale for concern:

LOW/HIGH/UNCLEAR

Y=YES, PY=PROBABLY YES, PN=PROBABLY NO, N=NO, NI=NO

DOMAIN 3: DATA COLLECTION AND STUDY APPRAISAL

Describe methods of data collection, what data were extracted from studies or collected through other means, how risk of bias was assessed (e.g. number of reviewers involved) and the tool used to assess risk of bias:

- | | |
|---|--------------|
| 1. Were efforts made to minimise error in data collection? | Y/PY/PN/N/NI |
| 2. Were sufficient study characteristics available for both review authors and readers to be able to interpret the results? | Y/PY/PN/N/NI |
| 3. Were all relevant study results collected for use in the synthesis? | Y/PY/PN/N/NI |
| 4. Was risk of bias (or methodological quality) formally assessed using appropriate criteria? | Y/PY/PN/N/NI |

Concerns regarding methods used to collect data and appraise studies

Rationale for concern:

LOW/HIGH/UNCLEAR

Y=YES, PY=PROBABLY YES, PN=PROBABLY NO, N=NO, NI=NO

DOMAIN 4: SYNTHESIS AND FINDINGS

Describe synthesis methods

1. Did the synthesis include all studies that it should?	Y/PY/PN/N/NI
2. Were all pre-defined analyses reported or departures explained?	Y/PY/PN/N/NI
3. Was the synthesis appropriate given the nature and similarity in the research questions, study designs and outcomes across included studies?	Y/PY/PN/N/NI
4. Was between study variation (heterogeneity) minimal or addressed in the synthesis?	Y/PY/PN/N/NI
5. Were the findings robust, e.g. as demonstrated through funnel plot or sensitivity analyses?	Y/PY/PN/N/NI

Concerns regarding the synthesis and findings

Rationale for concern:

LOW/HIGH/UNCLEAR

Y=YES, PY=PROBABLY YES, PN=PROBABLY NO, N=NO, NI=NO

10.3.3 Phase 3: Judging risk of bias

Summarise the concerns identified during the Phase 2 assessment:

Domain	Concern	Rationale for concern
Concerns regarding specification of study eligibility criteria		
Concerns regarding methods used to identify and/or select studies		
Concerns regarding methods used to collect data and appraise studies		
Concerns regarding the synthesis and findings		

RISK OF BIAS IN THE REVIEW

Describe whether conclusions were supported by the evidence

1. Did the interpretation of findings address all of the concerns identified in Domains 1 to 4?	Y/PY/PN/N/NI
2. Was the relevance of identified studies to the review's research question appropriately considered?	Y/PY/PN/N/NI
3. Did the reviewers avoid emphasizing results on the basis of their statistical significance?	Y/PY/PN/N/NI

Risk of bias in the review

Rationale for risk:

RISK

LOW/HIGH/UNCLEAR

Y=YES, PY=PROBABLY YES, PN=PROBABLY NO, N=NO, NI=NO



Suggested Tabular Presentation for ROBIS Results

Review	Phase 2					Phase 3
	1. STUDY ELIGIBILITY CRITERIA	2. IDENTIFICATION AND SELECTION OF STUDIES	3. DATA COLLECTION AND STUDY APPRAISAL	4. SYNTHESIS AND FINDINGS		RISK OF BIAS IN THE REVIEW
1	😊	😊	😊	😊		😞
2	😊	😊	😊	😊		😊
3	😞	😞	😊	😊		😞
4	😞	😞	😊	😊		😊
5	😞	?	😊	😊		😞
6	😞	?	😊	😊		😊
7	😞	😊	😊	😊		😞
8	😞	😊	😊	😊		😞
9	😞	😊	😊	😊		😞
10	😞	😊	😊	😞		😞
11	😊	😊	😊	😊		😊

😊 = low risk; 😞 = high risk; ? = unclear risk

Copied from [Resources | Bristol Medical School: Population Health Sciences | University of Bristol](#)

10.4 QUADAS-2 – diagnostic studies

QUADAS-2 is the version of QUADAS that is recommended for use in systematic reviews to evaluate the risk of bias and applicability of primary diagnostic accuracy studies (see [website](#)).¹⁷ The methodology team state on the website that in April 2025, a new version QUADAS-3 will soon be available.

QUADAS-2, consists of four key domains: patient selection, index test, reference standard, flow and timing. A full description is presented in 4.4.1.



10.5 ROB-1 – RCTs

Risk of bias (RoB-1) can be used as the basis for the evaluation.⁹⁶

Table 25 – Cochrane Collaboration’s risk of bias tool 1 for assessing risk of bias

Domain	Support for judgement	Review authors’ judgement
Selection bias		
Random sequence generation	Describe the method used to generate the allocation sequence in sufficient detail to allow an assessment of whether it should produce comparable groups	Selection bias (biased allocation to interventions) due to inadequate generation of a randomised sequence
Allocation concealment	Describe the method used to conceal the allocation sequence in sufficient detail to determine whether intervention allocations could have been foreseen in advance of, or during, enrolment	Selection bias (biased allocation to interventions) due to inadequate concealment of allocations prior to assignment
Performance bias		
Blinding of participants and personnel Assessments should be made for each main outcome (or class of outcomes)	Describe all measures used, if any, to blind study participants and personnel from knowledge of which intervention a participant received. Provide any information relating to whether the intended blinding was effective	Performance bias due to knowledge of the allocated interventions by participants and personnel during the study
Detection bias		
Blinding of outcome assessment Assessments should be made for each main outcome (or class of outcomes)	Describe all measures used, if any, to blind outcome assessors from knowledge of which intervention a participant received. Provide any information relating to whether the intended blinding was effective	Detection bias due to knowledge of the allocated interventions by outcome assessors
Attrition bias		
Incomplete outcome data Assessments should be made for each main outcome (or class of outcomes)	Describe the completeness of outcome data for each main outcome, including attrition and exclusions from the analysis. State whether attrition and exclusions were reported, the numbers in each intervention group (compared with total	Attrition bias due to amount, nature or handling of incomplete outcome data



	randomised participants), reasons for attrition/exclusions where reported, and any reinclusions in analyses performed by the review authors	
Reporting bias		
Selective reporting	State how the possibility of selective outcome reporting was examined by the review authors, and what was found	Reporting bias due to selective outcome reporting
Other bias		
Other sources of bias	State any important concerns about bias not addressed in the other domains in the tool If particular questions/entries were prespecified in the review's protocol, responses should be provided for each question/entry	Bias due to problems not covered elsewhere in the table

10.6 ROBINS-I – Non-randomised Studies of Interventions

Per outcome, seven domains need to be assessed: confounding bias, selection bias, bias in classification of interventions, bias due to deviations from intended intervention, bias due to missing data, bias in measurement outcomes and bias in selection of the reported result. A key feature of the tool is the inclusion of signalling questions within each domain of bias. The overall risk of bias judgement of a study outcome is based on the response options for each domain level. The risk of bias judgements are: low risk of bias (the study is comparable to a well-performed randomised controlled trial), moderate risk of bias (the study is sound for a non-randomised controlled trial but cannot be compared to a RCT), serious risk of bias (the study has some important problems), critical risk of bias (the study is too problematic to provide any useful evidence on the effects of the intervention) or no information to base a judgement. When at least one individual domain is considered at a serious risk of bias, the overall risk will become serious risk of bias, irrespective of other more 'positively judged' domains. In general, it rarely happens that a non-randomised controlled trial is judged as having a low risk of bias.



■ REFERENCES

1. Jespers V, Costa E, Ricour C, Verleye L. Klinische praktijkrichtlijn voor de diagnose, behandeling en follow up van cutaan melanoom voor de tweede lijn. Good Clinical Practice (GCP). Brussel. Federaal Kenniscentrum voor de Gezondheidszorg (KCE). 2025. KCE Reports 409A.
2. AGREE Next Steps Consortium 2017. The AGREE II Instrument [Electronic version]. Retrieved September 24th, 2024 from <http://www.agreetrust.org>.
3. Brouwers MC, Kho ME, Browman GP, Burgers JS, Cluzeau F, Feder G, et al. AGREE II: advancing guideline development, reporting and evaluation in health care. CMAJ. 2010;182(18):E839-42.
4. Klugar M, Lotfi T, Darzi AJ, Reinap M, Klugarová J, Kantorová L, et al. GRADE guidance 39: using GRADE-ADOLOPMENT to adopt, adapt or create contextualized recommendations from source guidelines and evidence syntheses. J Clin Epidemiol. 2024;174:111494.
5. Schünemann HJ, Wiercioch W, Brozek J, Etzeandía-Ikobaltzeta I, Mustafa RA, Manja V, et al. GRADE Evidence to Decision (EtD) frameworks for adoption, adaptation, and de novo development of trustworthy recommendations: GRADE-ADOLOPMENT. J Clin Epidemiol. 2017;81:101-10.
6. Schünemann HJ, Brennan S, Akl EA, Hultcrantz M, Alonso-Coello P, Xia J, et al. The development methods of official GRADE articles and requirements for claiming the use of GRADE - A statement by the GRADE guidance group. J Clin Epidemiol. 2023;159:79-84.
7. Schünemann HJ, Neumann I, Hultcrantz M, Brignardello-Petersen R, Zeng L, Murad MH, et al. GRADE guidance 35: update on rating imprecision for assessing contextualized certainty of evidence and making decisions. J Clin Epidemiol. 2022;150:225-42.
8. Schünemann HJ. Interpreting GRADE's levels of certainty or quality of the evidence: GRADE for statisticians, considering review information size or less emphasis on imprecision? J Clin Epidemiol. 2016;75:6-15.
9. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. Cochrane Handbook for Systematic Reviews of Interventions version 6.5 (updated August 2024). Cochrane, 2024. Available from www.cochrane.org/handbook.
10. Perry R, Whitmarsh A, Leach V, Davies P. A comparison of two assessment tools used in overviews of systematic reviews: ROBIS versus AMSTAR-2. Syst Rev. 2021;10(1):273.
11. Page MJ, Sterne JAC, Boutron I, Hróbjartsson A, Kirkham JJ, Li T, et al. ROB-ME: a tool for assessing risk of bias due to missing evidence in systematic reviews with meta-analysis. BMJ. 2023;383:e076754.
12. Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. Bmj. 2017;358:j4008.
13. Whiting P, Savović J, Higgins JP, Caldwell DM, Reeves BC, Shea B, et al. ROBIS: A new tool to assess risk of bias in systematic reviews was developed. J Clin Epidemiol. 2016;69:225-34.
14. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. Bmj. 2019;366:l4898.
15. Crocker TF, Lam N, Jordão M, Brundle C, Prescott M, Forster A, et al. Risk-of-bias assessment using Cochrane's revised tool for randomized trials (RoB 2) was useful but challenging and resource-intensive: observations from a systematic review. J Clin Epidemiol. 2023;161:39-45.
16. Wang Y, Keitz S, Briel M, Glasziou P, Brignardello-Petersen R, Siemieniuk RAC, et al. Development of ROBUST-RCT: Risk Of Bias instrument for Use in SysTematic reviews-for Randomised Controlled Trials. Bmj. 2025;388:e081199.
17. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529-36.

18. Reitsma JB, Rutjes A, Whiting P, Yang B, Leeflang MM, Bossuyt PM, et al. Chapter 8 Assessing risk of bias and applicability. *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy*. Version 2.0 (updated July 2023). Cochrane, 2023. Available from <https://training.cochrane.org/handbook-diagnostic-test-accuracy/current>.
19. Davenport CF, Leeflang MM, Takwoingi Y, Deeks JJ. Use of QUADAS-2. Lesson 6.3: Cochrane Collaboration DTA online learning materials. The Cochrane collaboration, September 2014. Videocast.
20. The Cochrane Collaboration. Review Manager (RevMan) version 5.4. 2020.
21. Yang B, Mallett S, Takwoingi Y, Davenport CF, Hyde CJ, Whiting PF, et al. QUADAS-C: A Tool for Assessing Risk of Bias in Comparative Diagnostic Accuracy Studies. *Annals of Internal Medicine*. 2021;174(11):1592-9.
22. Kaul T, Kellerhuis BE, Damen JAA, Schuit E, Jenniskens K, van Smeden M, et al. Methodological quality assessment tools for diagnosis and prognosis research: overview and guidance. *J Clin Epidemiol*. 2025;177:111609.
23. Lee J, Mulder F, Leeflang M, Wolff R, Whiting P, Bossuyt PM. QUAPAS: An Adaptation of the QUADAS-2 Tool to Assess Prognostic Accuracy Studies. *Ann Intern Med*. 2022;175(7):1010-8.
24. Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ*. 2016;355:i4919.
25. Higgins JPT, Morgan RL, Rooney AA, Taylor KW, Thayer KA, Silva RA, et al. A tool to assess risk of bias in non-randomized follow-up studies of exposure effects (ROBINS-E). *Environ Int*. 2024;186:108602.
26. Lunny C, Higgins JPT, White IR, Dias S, Hutton B, Wright JM, et al. Risk of Bias in Network Meta-Analysis (RoB NMA) tool. *BMJ*. 2025;388:e079839.
27. Andrews JC, Schünemann HJ, Oxman AD, Pottie K, Meerpohl JJ, Coello PA, et al. GRADE guidelines: 15. Going from evidence to recommendation-determinants of a recommendation's direction and strength. *J Clin Epidemiol*. 2013;66(7):726-35.
28. Guyatt G, Hultcrantz M, Agoritsas T, Iorio A, Vandvik PO, Montori VM. Why Core GRADE is needed: introduction to a new series in The BMJ. *BMJ*. 2025;389:e081902.
29. Dewidar O, Lotfi T, Langendam MW, Parmelli E, Saz Parkinson Z, Solo K, et al. Good or best practice statements: proposal for the operationalisation and implementation of GRADE guidance. *BMJ Evid Based Med*. 2023;28(3):189-96.
30. Guyatt G, Agoritsas T, Brignardello-Petersen R, Mustafa RA, Rylance J, Foroutan F, et al. Core GRADE 1: overview of the Core GRADE approach. *BMJ*. 2025;389:e081903.
31. Guyatt G, Oxman AD, Sultan S, Brozek J, Glasziou P, Alonso-Coello P, et al. GRADE guidelines: 11. Making an overall rating of confidence in effect estimates for a single outcome and for all outcomes. *J Clin Epidemiol*. 2013;66(2):151-7.
32. Neumann I, Brennan S, Meerpohl J, Davoli M, Coello PA, Akl EA, et al. Overview of the GRADE approach. In: Neumann I, Schunemann H, editors. *The GRADE Book version 1.0* (updated September 2024): The GRADE working Group; 2025.
33. Zeng L, Brignardello-Petersen R, Hultcrantz M, Siemieniuk RAC, Santesso N, Traversy G, et al. GRADE guidelines 32: GRADE offers guidance on choosing targets of GRADE certainty of evidence ratings. *J Clin Epidemiol*. 2021;137:163-75.
34. Hultcrantz M, Schunemann HJ, Mustafa RA, Rind DM, Murad MH, Mayer M, et al. GRADE Certainty Ratings: Thresholds Rather Than Categories of Contextualization (GRADE Guidance 41). *Ann Intern Med*. 2025.
35. Hultcrantz M, Rind D, Akl EA, Treweek S, Mustafa RA, Iorio A, et al. The GRADE Working Group clarifies the construct of certainty of evidence. *J Clin Epidemiol*. 2017;87:4-13.
36. Murad MH, Mustafa RA, Schunemann HJ, Sultan S, Santesso N. Rating the certainty in evidence in the absence of a single estimate of effect. *Evid Based Med*. 2017;22(3):85-7.



37. Guyatt G, Wang Y, Eachempati P, Iorio A, Murad MH, Hultcrantz M, et al. Core GRADE 4: rating certainty of evidence-risk of bias, publication bias, and reasons for rating up certainty. *BMJ*. 2025;389:e083864.
38. Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, et al. GRADE guidelines: 4. Rating the quality of evidence--study limitations (risk of bias). *J Clin Epidemiol*. 2011;64(4):407-15.
39. Guyatt G, Zhao Y, Mayer M, Briel M, Mustafa R, Izcovich A, et al. GRADE guidance 36: updates to GRADE's approach to addressing inconsistency. *J Clin Epidemiol*. 2023;158:70-83.
40. Guyatt G, Schandelmaier S, Brignardello-Petersen R, De Beer H, Prasad M, Murad MH, et al. Core GRADE 3: rating certainty of evidence-assessing inconsistency. *BMJ*. 2025;389:e081905.
41. Schandelmaier S, Briel M, Varadhan R, Schmid CH, Devasenapathy N, Hayward RA, et al. Development of the Instrument to assess the Credibility of Effect Modification Analyses (ICEMAN) in randomized controlled trials and meta-analyses. *CMAJ*. 2020;192(32):E901-E6.
42. Sousa-Pinto B, Neumann I, Vieira RJ, Bognanni A, Marques-Cruz M, Gil-Mata S, et al. Quantitative assessment of inconsistency in meta-analysis using decision thresholds with two new indices. *J Clin Epidemiol*. 2025;181:111725.
43. Izcovich A, Chu DK, Mustafa RA, Guyatt G, Brignardello-Petersen R. A guide and pragmatic considerations for applying GRADE to network meta-analysis. *BMJ*. 2023;381:e074495.
44. Brignardello-Petersen R, Bonner A, Alexander PE, Siemieniuk RA, Furukawa TA, Rochwerg B, et al. Advances in the GRADE approach to rate the certainty in estimates from a network meta-analysis. *J Clin Epidemiol*. 2018;93:36-44.
45. Puhan MA, Schunemann HJ, Murad MH, Li T, Brignardello-Petersen R, Singh JA, et al. A GRADE Working Group approach for rating the quality of treatment effect estimates from network meta-analysis. *BMJ*. 2014;349:g5630.
46. Guyatt G, Iorio A, De Beer H, Owen A, Agoritsas T, Murad MH, et al. Core GRADE 5: rating certainty of evidence-assessing indirectness. *BMJ*. 2025;389:e083865.
47. Guyatt G, Zeng L, Brignardello-Petersen R, Prasad M, De Beer H, Murad MH, et al. Core GRADE 2: choosing the target of certainty rating and assessing imprecision. *BMJ*. 2025;389:e081904.
48. Zhai C, Guyatt G. Fixed-effect and random-effects models in meta-analysis. *Chin Med J (Engl)*. 2024;137(1):1-4.
49. Morgano GP, Wiercioch W, Piovani D, Neumann I, Nieuwlaat R, Piggott T, et al. Defining decision thresholds for judgments on health benefits and harms using the grading of recommendations assessment, development, and evaluation (GRADE) evidence to decision (EtD) frameworks: a randomized methodological study (GRADE-THRESHOLD). *J Clin Epidemiol*. 2025;179:111639.
50. Cohen J. *Statistical power analysis for the behavioral sciences*. 2nd ed. Hillsdale NJ: Erlbaum; 1988.
51. Skoetz N, Goldkuhle M, van Dalen EC, Akl EA, Trivella M, Mustafa RA, et al. GRADE guidelines 27: how to calculate absolute effects for time-to-event outcomes in summary of findings tables and Evidence Profiles. *J Clin Epidemiol*. 2020;118:124-31.
52. Hunter KE, Webster AC, Page MJ, Willson M, McDonald S, Berber S, et al. *Searching clinical trials registers: guide for systematic reviewers*. *BMJ*. 2022;377:e068791.
53. Brignardello-Petersen R, Mustafa RA, Siemieniuk RAC, Murad MH, Agoritsas T, Izcovich A, et al. GRADE approach to rate the certainty from a network meta-analysis: addressing incoherence. *J Clin Epidemiol*. 2019;108:77-85.
54. Brignardello-Petersen R, Guyatt GH, Mustafa RA, Chu DK, Hultcrantz M, Schünemann HJ, et al. GRADE guidelines 33: Addressing imprecision in a network meta-analysis. *J Clin Epidemiol*. 2021;139:49-56.
55. Brignardello-Petersen R, Tomlinson G, Florez I, Rind DM, Chu D, Morgan R, et al. Grading of recommendations assessment, development, and evaluation concept article 5: addressing intransitivity in a network meta-analysis. *J Clin Epidemiol*. 2023;160:151-9.

56. Brignardello-Petersen R, Murad MH, Walter SD, McLeod S, Carrasco-Labra A, Rochwerf B, et al. GRADE approach to rate the certainty from a network meta-analysis: avoiding spurious judgments of imprecision in sparse networks. *J Clin Epidemiol.* 2019;105:60-7.
57. Brignardello-Petersen R, Izcovich A, Rochwerf B, Florez ID, Hazlewood G, Alhazanni W, et al. GRADE approach to drawing conclusions from a network meta-analysis using a partially contextualised framework. *BMJ.* 2020;371:m3907.
58. Yepes-Nunez JJ, Li SA, Guyatt G, Jack SM, Brozek JL, Beyene J, et al. Development of the summary of findings table for network meta-analysis. *J Clin Epidemiol.* 2019;115:1-13.
59. Hao Q, Gao Y, Zhao Y, Murad MH, Mustafa R, Ansari MT, et al. GRADE concept 6: a novel application of external indirect evidence into GRADE ratings of evidence certainty in network meta-analysis. *J Clin Epidemiol.* 2023;163:95-101.
60. McKenzie J, Brennan S. Chapter 12: Synthesizing and presenting findings using other methods [last updated October 2019]. In: JPT H, J T, J C, M C, T L, MJ P, et al., editors. *Cochrane Handbook for Systematic Reviews of Interventions Version 6.5: Cochrane;* 2024.
61. Brown LD, Cai TT, Dasgupta A. Interval Estimation for a Binomial Proportion. *Statistical Science.* 2001;16:101-17.
62. Campbell M, McKenzie JE, Sowden A, Katikireddi SV, Brennan SE, Ellis S, et al. Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline. *BMJ.* 2020;368:l6890.
63. Hsu J, Brozek JL, Terracciano L, Kreis J, Compalati E, Stein AT, et al. Application of GRADE: making evidence-based recommendations about diagnostic tests in clinical practice guidelines. *Implement Sci.* 2011;6:62.
64. Schünemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *Bmj.* 2008;336(7653):1106-10.
65. Schünemann HJ, Mustafa R, Brozek J, Santesso N, Alonso-Coello P, Guyatt G, et al. GRADE Guidelines: 16. GRADE evidence to decision frameworks for tests in clinical practice and public health. *J Clin Epidemiol.* 2016;76:89-98.
66. Yang B, Mustafa RA, Bossuyt PM, Brozek J, Hultcrantz M, Leeflang MMG, et al. GRADE Guidance: 31. Assessing the certainty across a body of evidence for comparative test accuracy. *J Clin Epidemiol.* 2021;136:146-56.
67. Schünemann HJ, Mustafa RA, Brozek J, Steingart KR, Leeflang M, Murad MH, et al. GRADE guidelines: 21 part 2. Test accuracy: inconsistency, imprecision, publication bias, and other domains for rating the certainty of evidence and presenting it in evidence profiles and summary of findings tables. *J Clin Epidemiol.* 2020;122:142-52.
68. Schünemann HJ, Mustafa RA, Brozek J, Steingart KR, Leeflang M, Murad MH, et al. GRADE guidelines: 21 part 1. Study design, risk of bias, and indirectness in rating the certainty across a body of evidence for test accuracy. *J Clin Epidemiol.* 2020;122:129-41.
69. Schünemann HJ, Mustafa RA, Brozek J, Santesso N, Bossuyt PM, Steingart KR, et al. GRADE guidelines: 22. The GRADE approach for tests and strategies-from test accuracy to patient-important outcomes and recommendations. *J Clin Epidemiol.* 2019;111:69-82.
70. Deeks JJ, Bossuyt PM, Leeflang MM, Takwoingi Y (editors). *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy. Version 2.0 (updated July 2023).* Cochrane, 2023. Available from <https://training.cochrane.org/handbook-diagnostic-test-accuracy/current>; 2023.
71. Schunemann H, Brozek J, Guyatt G, Oxman AD. Introduction to the GRADE Handbook. Chapter the GRADE approach for diagnostic tests and strategies. In: Neumann I, Schunemann H, editors. *The GRADE Book version 1.0 (updated September 2024): The GRADE working Group;* 2025.
72. Bossuyt PM. Chapter 3: Understanding the design of test accuracy studies. . Deeks JJ, Bossuyt PM, Leeflang MM, Takwoingi Y (editors). *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy. Version 2.0 (updated July 2023).* Cochrane, 2023. Available from <https://training.cochrane.org/handbook-diagnostic-test-accuracy/current>; 2023.



73. Harris RP, Helfand M, Woolf SH, Lohr KN, Mulrow CD, Teutsch SM, et al. REPRINT OF: Current Methods of the U.S. Preventive Services Task Force: A Review of the Process. *Am J Prev Med.* 2020;58(3):316-31.
74. Mustafa RA, El Mikati IK, Murad MH, Hultcrantz M, Steingart KR, Yang B, et al. GRADE guidance 37: rating imprecision in a body of evidence on test accuracy. *J Clin Epidemiol.* 2024;165:111189.
75. Korevaar DA, Gopalakrishna G, Cohen JF, Bossuyt PM. Targeted test evaluation: a framework for designing diagnostic accuracy studies with clear study hypotheses. *Diagn Progn Res.* 2019;3:22.
76. Struyf T, Boon HA, van de Pol AC, Tournoy J, Schuermans A, Verheij TJM, et al. Diagnosing serious infections in older adults presenting to ambulatory care: a systematic review. *Age Ageing.* 2021;50(2):405-14.
77. Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JA. A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics.* 2007;8(2):239-51.
78. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J Clin Epidemiol.* 2005;58(10):982-90.
79. Steinhauser S, Schumacher M, Rücker G. Modelling multiple thresholds in meta-analysis of diagnostic test accuracy studies. *BMC Med Res Methodol.* 2016;16(1):97.
80. Struyf T, Deeks JJ, Dinnes J, Takwoingi Y, Davenport C, Leeflang MM, et al. Signs and symptoms to determine if a patient presenting in primary care or hospital outpatient settings has COVID-19 disease. *Cochrane Database Syst Rev.* 2020;7(7):Cd013665.
81. Mustafa RA, Schünemann H. Challenges with need to improve the reference standard in diagnosis: paper 2: real-world case examples from GRADE for when a reference standard is not clearly defined/ easily replicable or when it is infeasible/ unethical to perform the reference standard. *J Clin Epidemiol.* 2023;154:206-7.
82. Takwoingi Y, Dendukuri N, Schiller I, Rücker G, Jones HE, Partlett C, et al. Chapter 10 Undertaking meta-analysis. In: Deeks JJ, Bossuyt PM, Leeflang MM, Takwoingi Y (editors). *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy.* Version 2.0 (updated July 2023). Cochrane. 2023.
83. Korevaar DA, van Es N, Zwinderman AH, Cohen JF, Bossuyt PM. Time to publication among completed diagnostic accuracy studies: associated with reported accuracy estimates. *BMC Med Res Methodol.* 2016;16:68.
84. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *J Clin Epidemiol.* 2005;58(9):882-93.
85. van Enst WA, Ochodo E, Scholten RJ, Hooft L, Leeflang MM. Investigation of publication bias in meta-analyses of diagnostic test accuracy: a meta-epidemiological study. *BMC Med Res Methodol.* 2014;14:70.
86. Brożek JL, Akl EA, Jaeschke R, Lang DM, Bossuyt P, Glasziou P, et al. Grading quality of evidence and strength of recommendations in clinical practice guidelines: Part 2 of 3. The GRADE approach to grading quality of evidence about diagnostic tests and strategies. *Allergy.* 2009;64(8):1109-16.
87. Guyatt G, Vandvik PO, Iorio A, Agarwal A, Yao L, Eachempati P, et al. Core GRADE 7: principles for moving from evidence to recommendations and decisions. *BMJ.* 2025;389:e083867.
88. Zhang Y, Alonso-Coello P, Guyatt GH, Yepes-Núñez JJ, Akl EA, Hazlewood G, et al. GRADE Guidelines: 19. Assessing the certainty of evidence in the importance of outcomes or values and preferences-Risk of bias and indirectness. *J Clin Epidemiol.* 2019;111:94-104.
89. Zhang Y, Coello PA, Guyatt GH, Yepes-Núñez JJ, Akl EA, Hazlewood G, et al. GRADE guidelines: 20. Assessing the certainty of evidence in the importance of outcomes or values and preferences-inconsistency, imprecision, and other domains. *J Clin Epidemiol.* 2019;111:83-93.



90. Xie F, Shemilt I, Vale L, Ruiz F, Drummond MF, Lord J, et al. GRADE guidance 23: considering cost-effectiveness evidence in moving from evidence to health-related recommendations. *J Clin Epidemiol*. 2023;162:135-44.
91. Pottie K, Welch V, Morton R, Akl EA, Eslava-Schmalbach JH, Katikireddi V, et al. GRADE equity guidelines 4: considering health equity in GRADE guideline development: evidence to decision process. *J Clin Epidemiol*. 2017;90:84-91.
92. Dewidar O, Darzi AJ, Sayfi S, Pardo JP, Welch V, Wright GC, et al. Seven principles for integrating health equity considerations in the practice guideline enterprise. *Journal of Clinical Epidemiology*. 2025;182:111777.
93. Dewidar O, Pardo JP, Welch V, Hazlewood GS, Darzi AJ, Barnabe C, et al. Operationalizing the GRADE-equity criterion to inform guideline recommendations: application to a medical cannabis guideline. *Journal of Clinical Epidemiology*. 2024;165:111185.
94. Andrews J, Guyatt G, Oxman AD, Alderson P, Dahm P, Falck-Ytter Y, et al. GRADE guidelines: 14. Going from evidence to recommendations: the significance and presentation of recommendations. *J Clin Epidemiol*. 2013;66(7):719-25.
95. Akl EA, Guyatt GH, Irani J, Feldstein D, Wasi P, Shaw E, et al. "Might" or "suggest"? No wording approach was clearly superior in conveying the strength of recommendation. *Journal of Clinical Epidemiology*. 2012;65(3):268-75.
96. Higgins JPT, Altman DG, on behalf of the Cochrane Statistical Methods Group and the Cochrane Bias Methods Group. Assessing risk of bias in included studies. In: Higgins JPT, Green S, editors. *Cochrane handbook for systematic reviews of interventions*: John Wiley & Sons; 2012. p. 187-241; 2012.