National Institute for Health and Care Excellence

Final

# Rheumatoid arthritis in adults: diagnosis and management

## Rheumatoid arthritis: methods

*NICE guideline NG100*

*Methods*

*July 2018*

*Final*

*Developed by the National Guideline Centre*

**Disclaimer**

The recommendations in this guideline represent the view of NICE, arrived at after careful consideration of the evidence available. When exercising their judgement, professionals are expected to take this guideline fully into account, alongside the individual needs, preferences and values of their patients or service users. The recommendations in this guideline are not mandatory and the guideline does not override the responsibility of healthcare professionals to make decisions appropriate to the circumstances of the individual patient, in consultation with the patient and, where appropriate, their carer or guardian.

Local commissioners and providers have a responsibility to enable the guideline to be applied when individual health professionals and their patients or service users wish to use it. They should do so in the context of local and national priorities for funding and developing services, and in light of their duties to have due regard to the need to eliminate unlawful discrimination, to advance equality of opportunity and to reduce health inequalities. Nothing in this guideline should be interpreted in a way that would be inconsistent with compliance with those duties.

NICE guidelines cover health and care in England. Decisions on how they apply in other UK countries are made by ministers in the Welsh Government, Scottish Government, and Northern Ireland Executive. All NICE guidance is subject to regular review and may be updated or withdrawn.


ISBN: 978-1-4731-3003-6

# Contents

# 1 Development of the guideline

## 1.1 What is a NICE guideline?

NICE guidelines are recommendations for the care of individuals in specific clinical conditions or circumstances within the NHS – from prevention and self-care through primary and secondary care to more specialised services. These may also include elements of social care or public health measures. We base our guidelines on the best available research evidence, with the aim of improving the quality of healthcare. We use predetermined and systematic methods to identify and evaluate the evidence relating to specific review questions.

NICE guidelines can:

- provide recommendations for the treatment and care of people by health professionals
- be used to develop standards to assess the clinical practice of individual health professionals
- be used in the education and training of health professionals
- help patients to make informed decisions
- improve communication between patient and health professional.

While guidelines assist the practice of healthcare professionals, they do not replace their knowledge and skills.

We produce our guidelines using the following steps:

- A guideline topic is referred to NICE from NHS England.
- Stakeholders register an interest in the guideline and are consulted throughout the development process.
- The scope is prepared by the National Guideline Centre (NGC).
- The NGC establishes a guideline committee.
- A draft guideline is produced after the group assesses the available evidence and makes recommendations.
- There is a consultation on the draft guideline.
- The final guideline is produced.

The guideline is made up of a collection of documents including this Methods report and a number of evidence reports covering each of the review questions included in the guideline. These can all be downloaded from NICE at www.nice.org.uk.

NICE also publishes a summary of the recommendation in this guideline, known as ‘the NICE guideline’.

NICE Pathways brings together all connected NICE guidance.

## 1.2 Remit

NICE received the remit for this guideline from NHS England. NICE commissioned the NGC to produce the guideline.

The remit for this guideline is:

to update the NICE guideline on rheumatoid arthritis in adults: management (CG79) as set out in the surveillance review decision.

ISBN: 978-1-4731-3003-6

## 1.3 Who developed this guideline?

A multidisciplinary guideline committee comprising health professionals and researchers as well as lay members developed this guideline (see the list of guideline committee members and the acknowledgements).

The National Institute for Health and Care Excellence (NICE) funds the National Guideline Centre (NGC) and thus supported the development of this guideline. The committee was convened by the NGC and chaired by Stephen Ward in accordance with guidance from NICE.

The group met approximately every 6 weeks during the development of the guideline. At the start of the guideline development process all committee members declared interests including consultancies, fee-paid work, shareholdings, fellowships and support from the healthcare industry. At all subsequent committee meetings, members declared arising conflicts of interest.

Members were either required to withdraw completely or for part of the discussion if their declared interest made it appropriate. The details of declared interests and the actions taken are shown in the declaration of interest register for this guideline published on the NICE website.

Staff from the NGC provided methodological support and guidance for the development process. The team working on the guideline included a project manager, systematic reviewers (research fellows), health economists and information specialists. They undertook systematic searches of the literature, appraised the evidence, conducted meta-analysis and cost-effectiveness analysis where appropriate and drafted the guideline in collaboration with the committee.

### 1.3.1 What this guideline covers

This guideline is a partial update of NICE guideline Rheumatoid arthritis in adults: management.[3] It updates a number of recommendations while also investigating clinical areas not addressed by the previous guideline. The population covered is adults (people who are 18 years old and older) with rheumatoid arthritis. The clinical areas are ultrasound for diagnosis and ongoing monitoring, prognostic factors, treatment with conventional disease modifying anti-rheumatic drugs (DMARDs), bridging treatment using glucocorticoids and analgesic treatment as well as frequency of monitoring and the treat to target approach.

For further details please refer to the scope for this guideline (published on the NICE website) and the review questions in section 2.1.

### 1.3.2 What this guideline does not cover

This guideline does not cover people with other causes of chronic inflammatory polyarthritis. Areas from Rheumatoid arthritis in adults: management.[3] that have not been updated are:

- Biological and targeted synthetic DMARDs for managing rheumatoid arthritis.
- Support for patients and carers in managing rheumatoid arthritis through education, self-management and the provision of information and advice.
- Location of review.
- Non-specialist referral to specialist services.
- Non-pharmacological treatments for managing rheumatoid arthritis, including:
    - podiatry
    - physiotherapy
    - occupational therapy
    - diet

ISBN: 978-1-4731-3003-6

    - complementary and alternative interventions or approaches.
- Multidisciplinary teams.
- Timing of referral for surgery.

### 1.3.3 Relationships between the guideline and other NICE guidance

**Related NICE technology appraisals:** [4]

- Certolizumab pegol for treating rheumatoid arthritis after inadequate response to a TNF-alpha inhibitor. NICE technology appraisal guidance 415 (2016)
- Adalimumab, etanercept, infliximab, certolizumab, pegol, golimumbab, tocilizumab and abatacept for rheumatoid arthritis not previously treated with DMARDs or after conventional DMARDs only have failed (2016). NICE technology appraisal guidance 375 (2016).
- Tocilizumab for the treatment of rheumatoid arthritis. NICE technology appraisal guidance 247 (2012)
- Golimumab for the treatment of rheumatoid arthritis after the failure of previous disease-modifying anti-rheumatic drugs. NICE technology appraisal guidance 225 (2011)
- Adalimumab, etanercept, infliximab, rituximab and abatacept for the treatment of rheumatoid arthritis after the failure of a TNF inhibitor. NICE technology appraisal guidance 195 (2010)

**Related NICE guidelines:** [6]

- Multimorbidity: clinical assessment and management. NICE guideline NG56 (2016)
- Cardiovascular disease: risk assessment and reduction, including lipid modification. NICE guideline CG181 (2015)
- Medicines optimisation: the safe and effective use of medicines to enable the best possible outcomes. NICE guideline NG5 (2015)
- Depression in adults with a chronic physical health problem: recognition and management. NICE guideline CG91 (2009)
- Osteoporosis: assessing the risk of fragility fracture. NICE guideline CG141 (2012)
- Medicines adherence. NICE guideline CG76 (2009)

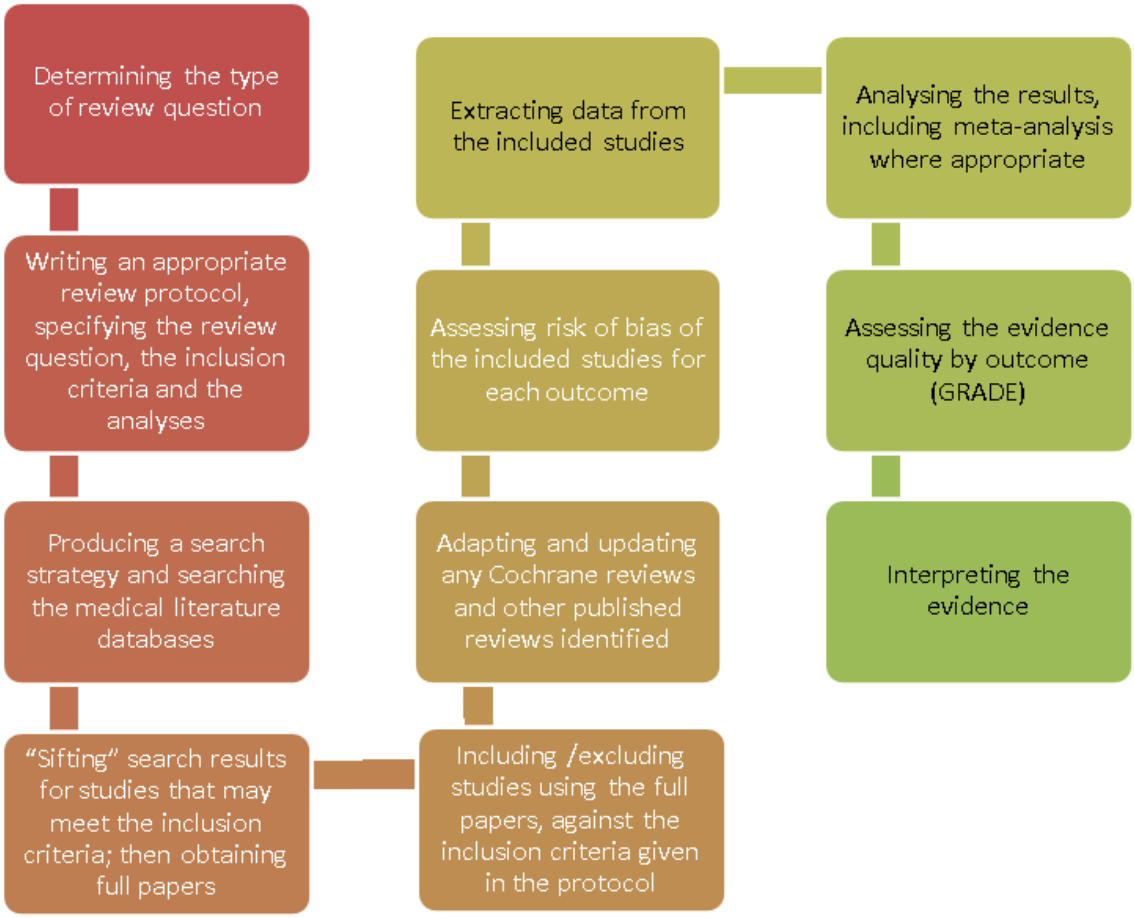
ISBN: 978-1-4731-3003-6

# 2 Methods

This report sets out in detail the methods used to review the evidence and to develop the recommendations that are presented in each of the evidence reviews for this guideline. This guidance was developed in accordance with the methods outlined in the NICE guidelines manual, 2014 version.[4]

Sections 2.1 to 2.3 describe the process used to identify and review clinical evidence (summarised in Figure 1), sections 2.2 and 2.4 describe the process used to identify and review the health economic evidence, and section 2.5 describes the process used to develop recommendations.

**Figure 1: Step-by-step process of review of evidence in the guideline**




## 2.1 Developing the review questions and outcomes

Review questions were developed using a PICO framework (population, intervention, comparison and outcome) for intervention reviews; using a framework of population, index tests, reference standard and target condition for reviews of diagnostic test accuracy; using population, presence or absence of factors under investigation (for example prognostic factors) and outcomes for prognostic reviews.

This use of a framework guided the literature searching process, critical appraisal and synthesis of evidence, and facilitated the development of recommendations by the guideline committee. The review questions were drafted by the NGC technical team and refined and validated by the committee. The questions were based on the key clinical areas identified in the scope.

ISBN: 978-1-4731-3003-6

A total of 14 review questions were identified. Three of these were separated into 2 strata to look at people with poor prognosis separately, leading to 17 reviews.

Full literature searches, critical appraisals and evidence reviews were completed for all the specified review questions.

**Table 1: Review questions**

| Evidence report | Type of review | Review questions | Outcomes |
|---|---|---|---|
| A | Diagnostic effectiveness and diagnostic accuracy | In adults with suspected inflammatory arthritis (including rheumatoid arthritis), what is the added value of ultrasound in the diagnosis of rheumatoid arthritis? | Critical clinical effectiveness outcomes:<br>• Disease Activity Score<br>• Quality of life<br>• Function<br><br>Important:<br>• Definitive clinical diagnosis<br>• Change/reclassification of diagnosis<br>• Change in management<br>• Number of prescribed DMARDs<br>• Number requiring repeat testing / additional testing.<br><br>Diagnostic accuracy outcomes:<br>• Sensitivity<br>• Specificity<br>• Positive predictive value<br>• Negative predictive value |
| B | Prognostic | In adults with rheumatoid arthritis, which risk factors are associated with poorer long-term function as measured by the Health Assessment Questionnaire (HAQ)? | Prognostic variables:<br>• HAQ scores at first presentation<br>• Elevated levels of CRP<br>• Elevated levels of ESR<br>• Presence or absence of RF<br>• Presence or absence of CCP or ACPA<br>• Presence or absence of X-ray erosion at first presentation<br>• Combinations of these factors (algorithm) |
| B | Prognostic | In adults with rheumatoid arthritis, which risk factors are associated with worse radiographic progression? | Prognostic variables:<br>• Elevated levels of CRP<br>• Elevated levels of ESR<br>• Presence or absence of RF<br>• Presence or absence of CCP or ACPA<br>• Presence or absence of X-ray erosion at first presentation |

ISBN: 978-1-4731-3003-6

| Evidence report | Type of review | Review questions | Outcomes |
|---|---|---|---|
| | | | • Combinations of these factors (algorithm) |
| C | Intervention | In adults with rheumatoid arthritis, what is the clinical and cost effectiveness of a treat-to-target management strategy, compared with usual care? | Critical:<br>• Disease Activity Score<br>• Quality of life<br>• Function<br><br>Important:<br>• Low disease activity<br>• Remission<br>• Fatigue<br>• Pain<br>• Radiological progression<br>• Withdrawal from trial / adherence to strategy |
| D | Intervention | In adults with rheumatoid arthritis, what is the best target to use when monitoring disease activity (remission or low disease activity)? | Critical:<br>• Disease Activity Score<br>• Quality of life<br>• Function<br><br>Important:<br>• Fatigue<br>• Pain<br>• Radiological progression<br>• Withdrawal / adherence |
| I | Intervention / Prognostic | In adults with rheumatoid arthritis, what is the added value of monitoring disease activity with ultrasound?<br><br>(People with poor prognosis will be reviewed as a separate strata). | Critical:<br>• Disease Activity Score<br>• Quality of life<br>• Function<br><br>Important:<br>• Remission<br>• Low disease activity<br>• Relapse<br>• Flare<br>• Pain<br>• Radiographic progression<br>• Change in planned management at time of testing<br>• Withdrawal from trial / adherence to strategy<br><br>Prognostic outcomes:<br>• Change in disease activity |
| E | Intervention | In adults with rheumatoid arthritis, what is the optimum frequency of disease activity monitoring (outside of the annual review)? | Critical:<br>• Disease Activity Score<br>• Quality of life<br>• Function |

ISBN: 978-1-4731-3003-6

| Evidence report | Type of review | Review questions | Outcomes |
|---|---|---|---|
| | | (People with poor prognosis will be reviewed as a separate strata). | Important:<br>• Remission<br>• Low disease activity<br>• Fatigue<br>• Pain<br>• Radiological progression<br>• Withdrawal due to adverse events<br>• Withdrawal due to inefficacy |
| G | Intervention | In adults with rheumatoid arthritis, what is the clinical and cost effectiveness of analgesics? | Critical:<br>• Pain<br>• Quality of life<br><br>Important:<br>• Stiffness<br>• Function<br>• Mortality<br>• Adverse events: Gastrointestinal (GI) effects<br>• Adverse events: Cardiac and vascular events<br>• Drug continuation |
| F | Intervention | In adults with RA who are DMARD naïve, which conventional DMARDs (alone or combined) are most clinically and cost effective?<br><br>(People with poor prognosis will be reviewed as a separate strata). | Critical:<br>• Disease Activity Score<br>• Quality of life<br>• Function<br><br>Important:<br>• Low disease activity<br>• Remission<br>• ACR50 response<br>• Pain<br>• Radiological progression<br>• Mortality<br>• Withdrawal due to adverse events<br>• Withdrawal due to inefficacy |
| F | Intervention | In adults with RA who are DMARD naïve, which DMARD treatment strategy (monotherapy, sequential monotherapy, parallel combination therapy, step up therapy or step down therapy) is most clinically and cost effective? | Critical:<br>• Disease Activity Score<br>• Quality of life<br>• Function<br><br>Important:<br>• Low disease activity<br>• Remission<br>• ACR50 response<br>• Pain<br>• Radiological progression |

ISBN: 978-1-4731-3003-6

| Evidence report | Type of review | Review questions | Outcomes |
|---|---|---|---|
| | | | • Mortality<br>• Withdrawal due to adverse events<br>• Withdrawal due to inefficacy |
| F | Intervention | In adults with RA who have had an inadequate response to, or failed treatment with, one or more conventional DMARDs, which conventional DMARDs (alone or combined) are most clinically and cost effective as subsequent treatment? | Critical:<br>• Disease Activity Score<br>• Quality of life<br>• Function<br><br>Important:<br>• Low disease activity<br>• Remission<br>• ACR50 response<br>• Pain<br>• Radiological progression<br>• Mortality<br>• Withdrawal due to adverse events<br>• Withdrawal due to inefficacy |
| F | Intervention | In adults with RA who have had an inadequate response to, or failed treatment with, one or more conventional DMARDs, which DMARD treatment strategy (monotherapy, sequential monotherapy, parallel combination therapy, step up therapy or step down therapy) is most clinically and cost effective as subsequent treatment? | Critical:<br>• Disease Activity Score<br>• Quality of life<br>• Function<br><br>Important:<br>• Low disease activity<br>• Remission<br>• ACR50 response<br>• Pain<br>• Radiological progression<br>• Mortality<br>• Withdrawal due to adverse events<br>• Withdrawal due to inefficacy |
| F | Intervention | In adults with rheumatoid arthritis, what is the clinical and cost effectiveness of adding short-term glucocorticoid (compared with placebo or no steroid treatment) when initiating a new DMARD? | Critical:<br>• Disease Activity Score<br>• Quality of life<br>• Function<br><br>Important:<br>• Low disease activity<br>• Remission<br>• Pain<br>• Continuing steroid use<br>• Radiological progression<br>• Adverse events (psychosis, hyperglycaemia, weight gain, insomnia, infection; dichotomous) |

ISBN: 978-1-4731-3003-6

| Evidence report | Type of review | Review questions | Outcomes |
|---|---|---|---|
| | | | • Withdrawal due to adverse events<br>• Withdrawal due to inefficacy |
| H | Intervention | In adults with rheumatoid arthritis, when initiating a new DMARD, which short-term glucocorticoid regime is most clinically and cost effective? | Critical:<br>• Disease Activity Score<br>• Quality of life<br>• Function<br><br>Important:<br>• Low disease activity<br>• Remission<br>• Pain<br>• Continuing steroid use<br>• Radiological progression<br>• Adverse events (psychosis, hyperglycaemia, weight gain, insomnia, infection; dichotomous)<br>• Withdrawal due to adverse events<br>• Withdrawal due to inefficacy |

## 2.2 Searching for evidence

**Clinical and health economics literature searches**

The full search strategy including population terms, intervention terms, study types applied, the databases searched and the years covered can be found in Appendix B of the evidence review report.

Systematic literature searches were undertaken to identify all published clinical and health economics evidence relevant to the review questions. Searches were undertaken according to the parameters stipulated within the NICE guidelines manual https://www.nice.org.uk/process/pmg20/. Databases were searched using relevant medical subject headings, free-text terms and study-type filters where appropriate. Studies published in languages other than English were not reviewed, where possible, searches were restricted to English Language. All searches were updated on 06 October 2017. Papers published or added to databases after this date were not considered. If new evidence falls outside of the timeframe for the guideline searches e.g. from stakeholder comments, the impact on the guideline will be considered, and any further action agreed between the developer and NICE staff with a quality assurance role.

Prior to running, searches were quality assured using different approaches. Medline search strategies were checked by a second information specialist before being run. Searches were cross-checked with reference lists of highly relevant papers, searches in other systematic reviews analysed, and committee members requested to highlight additional studies.

During the scoping stage, a search was conducted for guidelines and reports on the websites listed below. Web sites searched include:

- Guidelines International Network database (www.g-i-n.net)
- National Guideline Clearing House (www.guideline.gov)
- National Institute for Health and Care Excellence (NICE) (www.nice.org.uk)

ISBN: 978-1-4731-3003-6

- National Institutes of Health Consensus Development Program (consensus.nih.gov)
- NHS Evidence Search (www.evidence.nhs.uk).

Searching for unpublished literature was not undertaken. The NGC and NICE do not have access to drug manufacturers' unpublished clinical trial results, so the clinical evidence considered by the committee for pharmaceutical interventions may be different from that considered by the MHRA and European Medicines Agency for the purposes of licensing and safety regulation.

## 2.3 Identifying and analysing evidence of effectiveness

Research fellows conducted the tasks listed below, which are described in further detail in the rest of this section:

- Identified potentially relevant studies for each review question from the relevant search results by reviewing titles and abstracts. Full papers were then obtained.
- Reviewed full papers against prespecified inclusion and exclusion criteria to identify studies that addressed the review question in the appropriate population, and reported on outcomes of interest (review protocols are included in an appendix to each of the evidence reports).
- Critically appraised relevant studies using the appropriate study design checklist as specified in the NICE guidelines manual.[4] Prognostic studies were critically appraised using the amended QUIPS checklist.[2] Diagnostic accuracy studies were appraised using the Quality Assessment of Diagnostic Accuracy Studies version 2 (QUADAS-2) checklist. Intervention studies, including diagnostic randomised controlled trials, were appraised using 'Grading of Recommendations Assessment, Development and Evaluation (GRADE) toolbox' developed by the international GRADE working group (http://www.gradeworkinggroup.org/).
- Extracted key information about interventional study methods and results using 'Evibase', NGC's purpose-built software. Evibase produces summary evidence tables, including critical appraisal ratings. Key information about non-interventional study methods and results was manually extracted onto standard evidence tables and critically appraised separately (evidence tables are included in an appendix to each of the evidence reports).
- Generated summaries of the evidence by outcome. Outcome data were combined, analysed and reported according to study design:
  - Randomised data were meta-analysed where appropriate and reported in GRADE profile tables.
  - Prognostic data were meta-analysed where appropriate and reported in GRADE profile tables.
  - Diagnostic data meta-analysis would have been conducted where appropriate, that is, when 3 or more studies were available per threshold. However there was insufficient data to enable diagnostic meta-analysis.Data was presented as a range of values in adapted GRADE profile tables.
- A sample of a minimum of 10% of the abstract lists of the first 3 sifts by new reviewers and those for complex review questions (for example, prognostic reviews) were double-sifted by a senior research fellow and any discrepancies were rectified. All of the evidence reviews were quality assured by a senior research fellow. This included checking:
  - papers were included or excluded appropriately
  - a sample of the data extractions
  - correct methods were used to synthesise data
  - a sample of the risk of bias assessments.

ISBN: 978-1-4731-3003-6

### 2.3.1 Inclusion and exclusion criteria

The inclusion and exclusion of studies was based on the criteria defined in the review protocols, which can be found in an appendix to each of the evidence reports. Excluded studies (with the reasons for their exclusion) are listed in another appendix to each of the evidence reports. The committee was consulted about any uncertainty regarding inclusion or exclusion.

The key population inclusion criterion was:

- Adults with rheumatoid arthritis

Conference abstracts were not automatically excluded from any review. The abstracts were initially assessed against the inclusion criteria for the review question and further processed when a full publication was not available for that review question. If the abstracts were included the authors were contacted for further information. No relevant conference abstracts were identified for this guideline. Literature reviews, posters, letters, editorials, comment articles, unpublished studies and studies not in English were excluded.

### 2.3.2 Type of studies

Randomised trials, non-randomised intervention studies, and other observational studies (including diagnostic or prognostic studies) were included in the evidence reviews as appropriate.

For most intervention reviews in this guideline, parallel randomised controlled trials (RCTs) were included because they are considered the most robust type of study design that can produce an unbiased estimate of the intervention effects. Crossover RCTs were not deemed appropriate for any of the review questions. The committee agreed that in the majority of intervention reviews, lower quality evidence would not adequately inform changes in current practice, and thefore would not be included. One exception was made (steroids for bridging treatment) where the other study types included have been detailed in the protocol..

For the diagnostic review question, diagnostic RCTS, prospective cohort studies were included. For prognostic review questions, prospective cohort studies were included in the first instance. If insufficient were available then retrospective cohort studies were included. Case–control studies were not included.

### 2.3.3 Methods of combining clinical studies

#### 2.3.3.1 Data synthesis for intervention reviews

Where possible, meta-analyses were conducted using Cochrane Review Manager (RevMan5)[7] software to combine the data given in all studies for each of the outcomes of interest for the review question.

For some questions stratification was used for people with poor prognosis, this is documented in the individual review question protocols in each evidence report.

##### *2.3.3.1.1 Analysis of different types of data*

**Dichotomous outcomes**

Fixed-effects (Mantel–Haenszel) techniques (using an inverse variance method for pooling) were used to calculate risk ratios (relative risk, RR) for the binary outcomes, which included:

- mortality
- withdrawal due to inefficacy

ISBN: 978-1-4731-3003-6

- withdrawal due to adverse events
- low disease activity.
- remission
- ACR response.

The absolute risk difference was also calculated using GRADEpro[1] software, using the median event rate in the control arm of the pooled results.

For binary variables where there were zero events in either arm or a less than 1% event rate, Peto odds ratios, rather than risk ratios, were calculated. Peto odds ratios are more appropriate for data with a low number of events. Where Peto adds ratios have been used, risk difference utilised to calculate absolute effect.

For binary variables where there are zero events in both arms, the risk difference was utilised to calculate an absolute effect. Where sufficient information was provided, hazard ratios were calculated in preference for outcomes such as mortality where the time to the event occurring was important for decision-making.

**Continuous outcomes**

Continuous outcomes were analysed using an inverse variance method for pooling weighted mean differences. These outcomes included:

- heath-related quality of life (HRQoL)
- Disease Activity Score (DAS). DAS:0 to 10, DAS28: 0 to 9.4. Lower score indicates less disease activity.
- Function. HAQ: 0 to 3. Lower score indicates batter function.
- Pain. VAS: 0 to 100. Lower score indicates less pain.
- stiffness
- fatigue
- radiological progression. Sharp/van der Heijde score or Sharp Score. Range is dependent on number of joint analysed. Lower score is clinincally healthier.

Outcome timepoints of 1 year are often specified in the protocols to to reflect current process in terms of annual review for the person with RA with their rheumatologist.

Where the studies within a single meta-analysis had different scales of measurement, standardised mean differences were used (providing all studies reported either change from baseline or final values rather than a mixture of both); each different measure in each study was ‘normalised’ to the standard deviation value pooled between the intervention and comparator groups in that same study.

The means and standard deviations of continuous outcomes are required for meta-analysis. However, in cases where standard deviations were not reported, the standard error was calculated if the p values or 95% confidence intervals (95% CI) were reported, and meta-analysis was undertaken with the mean and standard error using the generic inverse variance method in Cochrane Review Manager (RevMan5[7] software. Where p values were reported as ‘less than’, a conservative approach was undertaken. For example, if a p value was reported as ‘p≤0.001’, the calculations for standard deviations were based on a p value of 0.001. If these statistical measures were not available then the methods described in section 16.1.3 of the Cochrane Handbook (version 5.1.0, updated March 2011) were applied.

#### *2.3.3.1.2 Generic inverse variance*

If a study reported only the summary statistic and 95% CI the generic-inverse variance method was used to enter data into RevMan5.[7] If the control event rate was reported this

ISBN: 978-1-4731-3003-6

was used to generate the absolute risk difference in GRADEpro.[1] If multivariate analysis was used to derive the summary statistic but no adjusted control event rate was reported no absolute risk difference was calculated.

#### *2.3.3.1.3 Heterogeneity*

Statistical heterogeneity was assessed for each meta-analysis estimate by considering the chi-squared test for significance at $p<0.1$ or an I-squared ($I^2$) inconsistency statistic (with an I-squared value of more than 50% indicating significant heterogeneity) as well as visual inspection of the distribution of effects. Where significant heterogeneity was present, predefined subgrouping of studies was carried out, this is documented in the individual review question protocols.

If the subgroup analysis resolved heterogeneity within all of the derived subgroups, then each of the derived subgroups were adopted as separate outcomes (providing at least 1 study remained in each subgroup). Assessments of potential differences in effect between subgroups were based on the chi-squared tests for heterogeneity statistics between subgroups. Any subgroup differences were interpreted with caution as separating the groups breaks the study randomisation and as such is subject to uncontrolled confounding.

Where heterogeneity was found, all subgrouping strategies were applied, the strategies were utilised independently, so subunits of subgroups were not created.

If all predefined strategies of subgrouping were unable to explain statistical heterogeneity within each derived subgroup, then a random effects (DerSimonian and Laird) model was employed to the entire group of studies in the meta-analysis. A random-effects model assumes a distribution of populations, rather than a single population. This leads to a widening of the confidence interval around the overall estimate, thus providing a more realistic interpretation of the true distribution of effects across more than 1 population. If, however, the committee considered the heterogeneity was so large that meta-analysis was inappropriate, then the results were described narratively.

#### *2.3.3.1.4 Complex analysis*

Network meta-analysis was considered for the comparison of first line cDMARD treatments, but was not pursued because it was not possible to create a connected network using any of the outcomes the committee prioritised (such as DAS, ACR50 response, DAS remission or DAS low disease activity).

### 2.3.3.2 Data synthesis for prognostic factor reviews

Odds ratios (ORs), risk ratios (RRs), or hazard ratios (HRs), with their 95% CIs, for the effect of the prespecified prognostic factors were extracted from the studies. Studies were only included if the confounders prespecified by the committee were either matched at baseline or were adjusted for in multivariate analysis. Matching at baseline was determined by the reported details of the people included in the trial.

Studies of lower risk of bias were preferred, taking into account the analysis and the study design. In particular, prospective cohort studies were preferred if they reported multivariable analyses that adjusted for key confounders identified by the committee at the protocol stage for that outcome.

Data were only combined in meta-analyses if the studies had adjusted for the same confounding factors and utilised similar populations and outcome measurements that could be combined. Where meta-analysis was not possible, each data point was presented separately in adapted GRADE profile tables.

ISBN: 978-1-4731-3003-6

#### 2.3.3.3 Data synthesis for diagnostic test accuracy reviews

The review protocols are separated into 2 sections to reflect the 2 different diagnostic study designs.

##### *2.3.3.3.1 Diagnostic RCTs*

Diagnostic RCTs (sometimes referred to as test and treat trials) are a randomised comparison of 2 diagnostic tests, with study outcomes being clinically important consequences of the diagnosis (patient-related outcome measures similar to those in intervention trials, such as mortality). Patients are randomised to receive test A or test B, followed by identical therapeutic interventions based on the results of the test (so someone with a positive result would receive the same treatment regardless of whether they were diagnosed by test A or test B). Downstream patient outcomes are then compared between the those assessed and treated via test A and those assessed and treated via test B. As treatment is the same in both arms of the trial, any differences in patient outcomes will reflect the accuracy of the tests in correctly establishing who does and does not have the condition. Data would be synthesised using the same methods for intervention reviews (see section 2.3.3.1.1 above).

##### *2.3.3.3.2 Diagnostic accuracy studies*

For diagnostic test accuracy studies, a positive result on the index test was found if the patient had values of the measured quantity above or below a threshold value. Different thresholds could be used and the thresholds were assessed by the committee including whether or not data could be pooled across a range of thresholds. Diagnostic test accuracy measures that were extracted where reported in relevant studies were: sensitivity, specificity, area under the receiver operating characteristics (ROC) curve (AUC), positive predictive value and negative predictive value. No calculations were undertaken to find measures where they were not reported. The threshold of a diagnostic test is defined as the value at which the test can best differentiate between those with and without the target condition. In practice this varies amongst studies. If a test has a high sensitivity then very few people with the condition will be missed (few false negatives). For example, a test with a sensitivity of 97% will only miss 3% of people with the condition. Conversely, if a test has a high specificity then few people without the condition would be incorrectly diagnosed (few false positives). For example, a test with a specificity of 97% will only incorrectly diagnose 3% of people who do not have the condition as positive. For this guideline, sensitivity was considered more important than specificity due to the consequences of a missed diagnosis (false negative result). A missed diagnosis can slow treatment and consequently lead to faster disease progression and increased joint damage**.**

Coupled forest plots of sensitivity and specificity with their 95% CIs across studies (at various thresholds) were produced where possible for each test, using RevMan5.[7] In order to do this, 2×2 tables (the number of true positives, false positives, true negatives and false negatives) were directly taken from the study if given, or else were derived from raw data or calculated from the set of test accuracy statistics.

Diagnostic meta-analysis would have been conducted where appropriate, that is, when 3 or more studies were available per threshold. However there was insufficient data to enable diagnostic meta-analysis.

The AUC describes the overall diagnostic accuracy across the full range of thresholds.

Heterogeneity or inconsistency amongst studies was visually inspected.

ISBN: 978-1-4731-3003-6

### 2.3.4 Appraising the quality of evidence by outcomes

#### 2.3.4.1 Intervention reviews

The evidence for outcomes from the included RCTs and, where appropriate, non-randomised intervention studies, were evaluated and presented using an adaptation of the GRADE toolbox developed by the international GRADE working group (http://www.gradeworkinggroup.org/). The software (GRADEpro[1]) developed by the GRADE working group was used to assess the quality of each outcome, taking into account individual study quality and the meta-analysis results.

Each outcome was first examined for each of the quality elements listed and defined in Table 2.

**Table 2: Description of quality elements in GRADE for intervention studies**

| Quality element | Description |
|---|---|
| Risk of bias | Limitations in the study design and implementation may bias the estimates of the treatment effect. Major limitations in studies decrease the confidence in the estimate of the effect. Examples of such limitations are selection bias (often due to poor allocation concealment), performance and detection bias (often due to a lack of blinding of the patient, healthcare professional or assessor) and attrition bias (due to missing data causing systematic bias in the analysis). |
| Indirectness | Indirectness refers to differences in study population, intervention, comparator and outcomes between the available evidence and the review question. |
| Inconsistency | Inconsistency refers to an unexplained heterogeneity of effect estimates between studies in the same meta-analysis. |
| Imprecision | Results are imprecise when studies include relatively few patients and few events (or highly variable measures) and thus have wide confidence intervals around the estimate of the effect relative to clinically important thresholds. 95% confidence intervals denote the possible range of locations of the true population effect at a 95% probability, and so wide confidence intervals may denote a result that is consistent with conflicting interpretations (for example a result may be consistent with both clinical benefit AND clinical harm) and thus be imprecise. |
| Publication bias | Publication bias is a systematic underestimate or overestimate of the underlying beneficial or harmful effect due to the selective publication of studies. A closely related phenomenon is where some papers fail to report an outcome that is inconclusive, thus leading to an overestimate of the effectiveness of that outcome. |
| Other issues | Sometimes randomisation may not adequately lead to group equivalence of confounders, and if so this may lead to bias, which should be taken into account. Potential conflicts of interest, often caused by excessive pharmaceutical company involvement in the publication of a study, should also be noted. |

Details of how the 4 main quality elements (risk of bias, indirectness, inconsistency and imprecision) were appraised for each outcome are given below. Publication or other bias was only taken into consideration in the quality assessment if it was apparent.

##### *2.3.4.1.1 Risk of bias*

The main domains of bias for RCTs are listed in Table 3. Each outcome had its risk of bias assessed within each study first. For each study, if there were no risks of bias in any domain, the risk of bias was given a rating of 0. If there was risk of bias in just 1 domain, the risk of bias was given a 'serious' rating of −1, but if there was risk of bias in 2 or more domains the risk of bias was given a 'very serious' rating of −2. A weighted average score was then calculated across all studies contributing to the outcome, by taking into account the weighting of studies according to study precision. For example if the most precise studies tended to

ISBN: 978-1-4731-3003-6

each have a score of −1 for that outcome, the overall score for that outcome would tend towards −1.

**Table 3: Principle domains of bias in randomised controlled trials**

| Limitation | Explanation |
|---|---|
| Selection bias (sequence generation and allocation concealment) | If those enrolling patients are aware of the group to which the next enrolled patient will be allocated, either because of a non-random sequence that is predictable, or because a truly random sequence was not concealed from the researcher, this may translate into systematic selection bias. This may occur if the researcher chooses not to recruit a participant into that specific group because of:<br>• knowledge of that participant's likely prognostic characteristics, and<br>• a desire for one group to do better than the other. |
| Performance and detection bias (lack of blinding of patients and healthcare professionals) | Patients, caregivers, those adjudicating or recording outcomes, and data analysts should not be aware of the arm to which patients are allocated. Knowledge of the group can influence:<br>• the experience of the placebo effect<br>• performance in outcome measures<br>• the level of care and attention received, and<br>• the methods of measurement or analysis<br>all of which can contribute to systematic bias. |
| Attrition bias | Attrition bias results from an unaccounted for loss of data beyond a certain level (a differential of 10% between groups). Loss of data can occur when participants are compulsorily withdrawn from a group by the researchers (for example, when a per-protocol approach is used) or when participants do not attend assessment sessions. If the missing data are likely to be different from the data of those remaining in the groups, and there is a differential rate of such missing data from groups, systematic attrition bias may result. |
| Selective outcome reporting | Reporting of some outcomes and not others on the basis of the results can also lead to bias, as this may distort the overall impression of efficacy. |
| Other limitations | For example:<br>• Stopping early for benefit observed in randomised trials, in particular in the absence of adequate stopping rules.<br>• Use of unvalidated patient-reported outcome measures.<br>• Lack of washout periods to avoid carry-over effects in crossover trials.<br>• Recruitment bias in cluster-randomised trials. |

#### *2.3.4.1.2 Indirectness*

Indirectness refers to the extent to which the populations, interventions, comparisons and outcome measures are dissimilar to those defined in the inclusion criteria for the reviews. Indirectness is important when these differences are expected to contribute to a difference in effect size, or may affect the balance of harms and benefits considered for an intervention. As for the risk of bias, each outcome had its indirectness assessed within each study first. For each study, if there were no sources of indirectness, indirectness was given a rating of 0. If there was indirectness in just 1 source (for example in terms of population), indirectness was given a 'serious' rating of −1, but if there was indirectness in 2 or more sources (for example, in terms of population and treatment) the indirectness was given a 'very serious' rating of −2. A weighted average score was then calculated across all studies contributing to the outcome by taking into account study precision. For example, if the most precise studies tended to have an indirectness score of −1 each for that outcome, the overall score for that outcome would tend towards −1.

ISBN: 978-1-4731-3003-6

#### *2.3.4.1.3 Inconsistency*

Inconsistency refers to an unexplained heterogeneity of results for an outcome across different studies. When estimates of the treatment effect across studies differ widely, this suggests true differences in the underlying treatment effect, which may be due to differences in populations, settings or doses. Statistical tests for heterogeneity, as detailed below, are heuristic and reviewers took the $I^2$ and chi-squared results into account alongside other information including the relative positions of study confidence intervals and point estimates with regards to the MIDs.

When heterogeneity existed within an outcome (chi-squared $p<0.1$, or $I^2>50\%$), but no plausible explanation could be found, the quality of evidence for that outcome was downgraded. Inconsistency for that outcome was given a 'serious' score of −1 if the $I^2$ was 50–74%, and a 'very serious' score of −2 if the $I^2$ was 75% or more.

If inconsistency could be explained based on prespecified subgroup analysis (that is, each subgroup had an $I^2<50\%$), the committee took this into account and considered whether to make separate recommendations on new outcomes based on the subgroups defined by the assumed explanatory factors. In such a situation the quality of evidence was not downgraded for those emergent outcomes.

Since the inconsistency score was based on the meta-analysis results, the score represented the whole outcome and so weighted averaging across studies was not necessary.

#### *2.3.4.1.4 Imprecision*

The criteria applied for imprecision were based on the 95% CIs for the pooled estimate of effect, and the minimal important differences (MID) for the outcome. The MIDs are the threshold for appreciable benefits and harms, separated by a zone either side of the line of no effect where there is assumed to be no clinically important effect. If either end of the 95% CI of the overall estimate of effect crossed 1 of the MID lines, imprecision was regarded as serious and a 'serious' score of −1 was given. This was because the overall result, as represented by the span of the confidence interval, was consistent with 2 interpretations as defined by the MID (for example, both no clinically important effect and clinical benefit were possible interpretations). If both MID lines were crossed by either or both ends of the 95% CI then imprecision was regarded as very serious and a 'very serious' score of −2 was given. This was because the overall result was consistent with all 3 interpretations defined by the MID (no clinically important effect, clinical benefit and clinical harm). This is illustrated in Figure 2. As for inconsistency, since the imprecision score was based on the meta-analysis results, the score represented the whole outcome and so weighted averaging across studies was not necessary.

The position of the MID lines is ideally determined by values reported in the literature. 'Anchor-based' methods aim to establish clinically meaningful changes in a continuous outcome variable by relating or 'anchoring' them to patient-centred measures of clinical effectiveness that could be regarded as gold standards with a high level of face validity. For example, a MID for an outcome could be defined by the minimum amount of change in that outcome necessary to make patients feel their quality of life had 'significantly improved'. MIDs in the literature may also be based on expert clinician or consensus opinion concerning the minimum amount of change in a variable deemed to affect quality of life or health. For binary variables, any MIDs reported in the literature will inevitably be based on expert consensus, as such MIDs relate to all-or-nothing population effects rather than measurable effects on an individual, and so are not amenable to patient-centred 'anchor' methods.

In the absence of values identified in the literature, the alternative approach to deciding on MID levels was as follows:

ISBN: 978-1-4731-3003-6

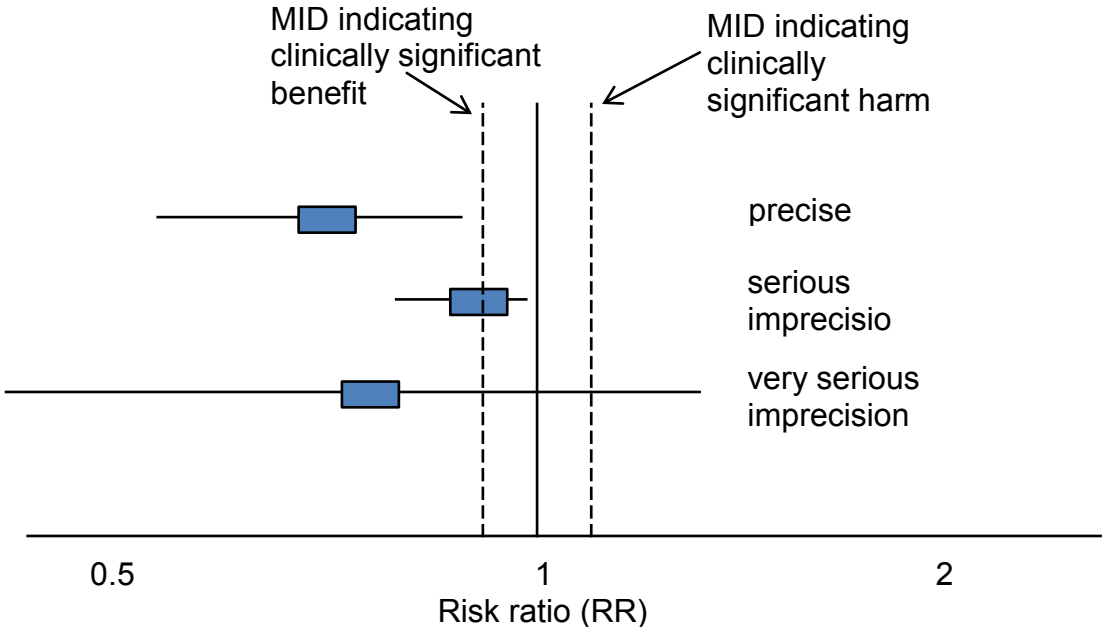
- For categorical outcomes the MIDs were taken to be RRs of 0.75 and 1.25. For 'positive' outcomes such as 'patient satisfaction', the RR of 0.75 is taken as the line denoting the boundary between no clinically important effect and a clinically important harm, whilst the RR of 1.25 is taken as the line denoting the boundary between no clinically important effect and a clinically important benefit. For 'negative' outcomes such as 'bleeding', the opposite occurs, so the RR of 0.75 is taken as the line denoting the boundary between no clinically important effect and a clinically important benefit, whilst the RR of 1.25 is taken as the line denoting the boundary between no clinically important effect and a clinically important harm.
- For continuous outcome variables the MID was taken as half the median baseline standard deviation of that variable, across all studies in the meta-analysis. Hence the MID denoting the minimum clinically important benefit was positive for a 'positive' outcome (for example, a quality of life measure where a higher score denotes better health), and negative for a 'negative' outcome (for example, a visual analogue scale [VAS] pain score). Clinically important harms will be the converse of these. If baseline values are unavailable, then half the median comparator group standard deviation of that variable will be taken as the MID.
- If standardised mean differences have been used, then the MID will be set at the absolute value of +0.5. This follows because standardised mean differences are mean differences normalised to the pooled standard deviation of the 2 groups, and are thus effectively expressed in units of 'numbers of standard deviations'. The 0.5 MID value in this context therefore indicates half a standard deviation, the same definition of MID as used for non-standardised mean differences.

The default MID value was subject to amendment after discussion with the committee. If the committee decided that the MID level should be altered, after consideration of absolute as well as relative effects, this was allowed, provided that any such decision was not influenced by any bias towards making stronger or weaker recommendations for specific outcomes. Peto odds ratios utilised the same process for imprecision as Mantel-Haenszel odds ratios.

For this guideline, imprecision was judged using the default method. A number of consensus MIDs were utilised for the judgement of clinical importance with discussion with the committee.

ISBN: 978-1-4731-3003-6

**Figure 2:** Illustration of precise and imprecise outcomes based on the 95% CI of dichotomous outcomes in a forest plot (Note that all 3 results would be pooled estimates, and would not, in practice, be placed on the same forest plot)




#### *2.3.4.1.5 Overall grading of the quality of clinical evidence*

Once an outcome had been appraised for the main quality elements, as above, an overall quality grade was calculated for that outcome. The scores (0, −1 or −2) from each of the main quality elements were summed to give a score that could be anything from 0 (the best possible) to −8 (the worst possible). However scores were capped at −3. This final score was then applied to the starting grade that had originally been applied to the outcome by default, based on study design. All RCTs started as High and the overall quality became Moderate, Low or Very Low if the overall score was −1, −2 or −3 points respectively. The significance of these overall ratings is explained in Table 4. The reasons for downgrading in each case were specified in the footnotes of the GRADE tables.

Non-randomised intervention studies started at Low, and so a score of −1 would be enough to take the grade to the lowest level of Very Low. Non-randomised intervention studies could, however, be upgraded if there was a large magnitude of effect or a dose-response gradient.

**Table 4: Overall quality of outcome evidence in GRADE**

| Level | Description |
|---|---|
| High | Further research is very unlikely to change our confidence in the estimate of effect |
| Moderate | Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate |
| Low | Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate |
| Very low | Any estimate of effect is very uncertain |

ISBN: 978-1-4731-3003-6

#### 2.3.4.2 Prognostic reviews

A modified GRADE methodology was used for prognostic studies, considering risk of bias, indirectness, inconsistency and imprecision. The data was presented in a modified GRADE table for each outcome and the prognostic data extracted from each study for the factors of interest utilised a line of the table.

#### 1.1.2.1 Risk of bias

The quality of evidence for prognostic studies was evaluated according to the criteria given in Table 5. If data were meta-analysed, the quality for pooled studies was presented. If the data were not pooled, then a quality rating was presented for each study. The criteria in Table 5 are taken from the amended Quality in Prognosis Study (QUIPS) tool.[2]

**Table 5: Description of quality elements for prospective studies**

| Quality element | Description of cases where the quality measure would be downgraded |
|---|---|
| Study Participation | Adequate source of population, adequately described inclusion/exclusion criteria, recruitment method clearly described, table of baseline factors reported. |
| Study attrition | Response rate, reasons for loss to follow up, no important differences between key characteristics and outcomes in study participants who completed the studies and those who did not |
| Prognostic factor measurement | Prognostic factor measurement clearly defined, data collection procedure adequate, any incomplete data taken into account for in the analysis, method/setting of measurement consistent across included studies |
| Outcome measurement | Clear definition of outcome, outcome measurement valid, method of measuring outcome consistent across included studies |
| Study confounding | All important confounders considered and measured, clear definition, adequate measurement of confounders, method of confounding measurement is consistent across included studies, appropriate imputation techniques applied for missing data if used, important potential confounders accounted for in the analysis |
| Statistical Analysis and Reporting | No selective reporting of results, analysis addressed missing data if appropriate, appropriate strategy for model building, selected model was adequate for the design of the review, including taking account of the time-to-event nature of the data |
| Other risk of bias | For example concerns that (retrospective or prospective) design caused risk of bias issues additional to those covered by the other domains |

#### 1.1.2.2 Indirectness

For each paper, if there were no sources of indirectness, indirectness was given a rating of 0. If there was indirectness in just one source (for example in terms of population), indirectness was given a 'serious' rating of -1, but if there was indirectness in two or more sources (for example, in terms of population and risk factor) the indirectness was given a 'very serious' rating of -2. A weighted average score was then calculated across all studies contributing to the outcome, by taking into account the weights in the meta-analysis.

##### *2.3.4.2.1 Inconsistency*

Inconsistency was assessed as for intervention studies.

ISBN: 978-1-4731-3003-6

#### *2.3.4.2.2 Imprecision*

The position of the 95% CIs in relation to the null line determined the existence of imprecision. If the 95% CI did not cross the null line then no serious imprecision was recorded. If the 95% CI crossed the null line then serious imprecision was recorded.

#### *2.3.4.2.3 Overall grading*

Because prognostic reviews were not usually based on multiple outcomes per study, quality rating was assigned by study. However if there was more than 1 outcome involved in a study, then the quality rating of the evidence statements for each outcome was adjusted accordingly. For example, if one outcome was based on an invalidated measurement method, but another outcome in the same study was not, the second outcome would be graded 1 grade higher than the first outcome.

Quality rating started at high for prospective cohort studies, and each major limitation brought the rating down by 1 increment to a minimum grade of very low, as explained for interventional reviews. For prognostic reviews prospective cohort studies with a multivariate analysis are regarded as the gold standard because RCTs are usually inappropriate for these types of review for ethical or pragmatic reasons. Furthermore, if the study is looking at more than 1 risk factor of interest then randomisation would be inappropriate as it can only be applied to 1 of the risk factors.

### 2.3.4.3 Diagnostic studies

A modified GRADE methodology was used for diagnostic accuracy studies, considering risk of bias, indirectness, inconsistency and imprecision.

#### *1.1.2.2.1 Risk of bias*

Risk of bias and indirectness of evidence for diagnostic data were evaluated by study using the Quality Assessment of Diagnostic Accuracy Studies version 2 (QUADAS-2) checklists (see appendix H in the NICE guidelines manual 2014[4]). Risk of bias and applicability in primary diagnostic accuracy studies in QUADAS-2 consists of 4 domains (see Figure 3):

- patient selection
- index test
- reference standard
- flow and timing.

**Figure 3: Summary of QUADAS-2 with list of signalling, risk of bias and applicability questions.**

| Domain | Patient selection | Index test | Reference standard | Flow and timing |
|---|---|---|---|---|
| Description | Describe methods of patient selection. Describe included patients (prior testing, presentation, intended use of index test and setting) | Describe the index test and how it was conducted and interpreted | Describe the reference standard and how it was conducted and interpreted | Describe any patients who did not receive the index test(s) and/or reference standard or who were excluded from the 2×2 table (refer to flow diagram). Describe the time interval and any interventions between index test(s) and reference standard |
| Signalling questions | Was a consecutive or | Were the index test results | Is the reference standard likely to | Was there an appropriate interval |

ISBN: 978-1-4731-3003-6

| Domain | Patient selection | Index test | Reference standard | Flow and timing |
|---|---|---|---|---|
| (yes/no/ unclear) | random sample of patients enrolled? | interpreted without knowledge of the results of the reference standard? | correctly classify the target condition? | between index test(s) and reference standard? |
| | Was a case–control design avoided? | If a threshold was used, was it pre-specified? | Were the reference standard results interpreted without knowledge of the results of the index test? | Did all patients receive a reference standard? |
| | Did the study avoid inappropriate exclusions? | | | Did all patients receive the same reference standard? |
| | | | | Were all patients included in the analysis? |
| Risk of bias; (high/low/ unclear) | Could the selection of patients have introduced bias? | Could the conduct or interpretation of the index test have introduced bias? | Could the reference standard, its conduct or its interpretation have introduced bias? | Could the patient flow have introduced bias? |
| Concerns regarding applicability (high/low/ unclear) | Are there concerns that the included patients do not match the review question? | Are there concerns that the index test, its conduct, or interpretation differ from the review question? | Are there concerns that the target condition as defined by the reference standard does not match the review question? | |

#### *2.3.4.3.1 Inconsistency*

Inconsistency refers to an unexplained heterogeneity of results for an outcome across different studies. Inconsistency was assessed by inspection of the accuracy data value (based on the primary measure) using the point estimates and, where available, 95% CIs of the individual studies on the forest plots. Particular attention was placed on values above or below 50% (diagnosis based on chance alone) and the 90% sensitivity threshold set by the committee (the threshold above which it would be acceptable to recommend a test). Studies where inspection of the sensitivity varied from below 50% to above 90% were downgraded by 2 increments and studies where data bisected 50% or 90% were downgraded by 1 increment.

#### *2.3.4.3.2 Imprecision*

Imprecision was assessed based on inspection of the confidence region for sensitivity in the diagnostic analysis. The evidence was downgraded by 1 increment when there was a 20 to 40% range of the confidence interval around the point estimate, and downgraded by 2 increments when there was a range of over 40%.

#### *2.3.4.3.3 Overall grading*

Quality rating started at high and each major limitation (risk of bias, indirectness, inconsistency and imprecision) brought the rating down by 1 increment to a minimum grade of Very Low, as explained for intervention reviews.

ISBN: 978-1-4731-3003-6

### 2.3.5 Assessing clinical importance

The committee assessed the evidence by outcome in order to determine if there was, or potentially was, a clinically important benefit, a clinically important harm or no clinically important difference between interventions. To facilitate this, binary outcomes were converted into absolute risk differences (ARDs) using GRADEpro[1] software: the median control group risk across studies was used to calculate the ARD and its 95% CI from the pooled risk ratio.

The assessment of clinical benefit, harm, or no benefit or harm was based on the point estimate of absolute effect for intervention studies, which was standardised across the reviews. For some outcomes MIDs were applied through committee consensus:

- Disease Activity Score: a change of 0.6.
- Function: a HAQ change of 0.1.
- Pain: a change of 10 on a hundred point scale.
- Any change in radiological progression.
- Duration of stiffness: a change of 30 minutes.

For other outcomes, no appropriate MIDs outcomes were agreed, and so the default method was adopted.

- For categorical outcomes the MIDs were taken to be RRs of 0.75 and 1.25. For 'positive' outcomes such as 'patient satisfaction', the RR of 0.75 is taken as the line denoting the boundary between no clinically important effect and a clinically important harm, whilst the RR of 1.25 is taken as the line denoting the boundary between no clinically important effect and a clinically important benefit. For 'negative' outcomes such as 'bleeding', the opposite occurs, so the RR of 0.75 is taken as the line denoting the boundary between no clinically important effect and a clinically important benefit, whilst the RR of 1.25 is taken as the line denoting the boundary between no clinically important effect and a clinically important harm.
- For continuous outcome variables the MID was taken as half the median baseline standard deviation of that variable, across all studies in the meta-analysis. Hence the MID denoting the minimum clinically important benefit was positive for a 'positive' outcome (for example, a quality of life measure where a higher score denotes better health), and negative for a 'negative' outcome (for example, a visual analogue scale [VAS] pain score). Clinically important harms will be the converse of these. If baseline values are unavailable, then half the median comparator group standard deviation of that variable will be taken as the MID.
- If standardised mean differences have been used, then the MID will be set at the absolute value of +0.5. This follows because standardised mean differences are mean differences normalised to the pooled standard deviation of the 2 groups, and are thus effectively expressed in units of 'numbers of standard deviations'. The 0.5 MID value in this context therefore indicates half a standard deviation, the same definition of MID as used for non-standardised mean differences.

This assessment was carried out by the committee for each critical outcome, and evidence statements formulated to compile the committee's assessments of clinical importance per outcome, alongside the evidence quality.

Clinical importance of diagnostic accuracy outcomes was discussed for each test with the committee. The committee stated in the protocol whether sensitivity or specificity would be their primary outcome for decision making. In the case of the review for ultrasound for diagnosis, the committee agreed sensitivity was the priority outcomeand agreed a minimum threshold of 90% for recommending the test.

ISBN: 978-1-4731-3003-6

Prognostic data indicated whether prognoctos variables independently predicted the outcome of interest. Therefore there was no need to assign clinical importance to these results.

### 2.3.6 Clinical evidence statements

Clinical evidence statements are summary statements that are included in each evidence report, and which summarise the key features of the clinical effectiveness evidence presented. The wording of the evidence statements reflects the certainty or uncertainty in the estimate of effect. The evidence statements are presented by outcome and encompass the following key features of the evidence:

- The number of studies and the number of participants for a particular outcome.
- An indication of the direction of clinical importance (if one treatment is beneficial or harmful compared to the other, or whether there is no difference between the 2 tested treatments).
- A description of the overall quality of the evidence (GRADE overall quality). Where GRADE was not formally used, a modified GRADE approach was utilised in the formulation of evidence statements.

## 2.4 Identifying and analysing evidence of cost effectiveness

The committee is required to make decisions based on the best available evidence of both clinical effectiveness and cost effectiveness. Guideline recommendations should be based on the expected costs of the different options in relation to their expected health benefits (that is, their 'cost effectiveness') rather than the total implementation cost. However, the committee will also need to be increasingly confident in the cost effectiveness of a recommendation as the cost of implementation increases. Therefore, the committee may require more robust evidence on the effectiveness and cost effectiveness of any recommendations that are expected to have a substantial impact on resources; any uncertainties must be offset by a compelling argument in favour of the recommendation. The cost impact or savings potential of a recommendation should not be the sole reason for the committee's decision.[4]

Health economic evidence was sought relating to the key clinical issues being addressed in the guideline. Health economists:

- Undertook a systematic review of the published economic literature.

### 2.4.1 Literature review

The health economists:

- Identified potentially relevant studies for each review question from the health economic search results by reviewing titles and abstracts. Full papers were then obtained.
- Reviewed full papers against prespecified inclusion and exclusion criteria to identify relevant studies (see below for details).
- Critically appraised relevant studies using economic evaluations checklists as specified in the NICE guidelines manual.[4]
- Extracted key information about the studies' methods and results into health economic evidence tables (which can be found in appendices to the relevant evidence reports).
- Generated summaries of the evidence in NICE health economic evidence profile tables (included in the relevant evidence report for each review question) – see below for details.

ISBN: 978-1-4731-3003-6

#### 2.4.1.1 Inclusion and exclusion criteria

Full economic evaluations (studies comparing costs and health consequences of alternative courses of action: cost–utility, cost-effectiveness, cost–benefit and cost–consequences analyses) and comparative costing studies that addressed the review question in the relevant population were considered potentially includable as health economic evidence.

Studies that only reported cost per hospital (not per patient), or only reported average cost effectiveness without disaggregated costs and effects were excluded. Literature reviews, abstracts, posters, letters, editorials, comment articles, unpublished studies and studies not in English were excluded. Studies published before 2001 and studies from non-OECD countries or the USA were also excluded, on the basis that the applicability of such studies to the present UK NHS context is likely to be too low for them to be helpful for decision-making.

Remaining health economic studies were prioritised for inclusion based on their relative applicability to the development of this guideline and the study limitations. For example, if a high quality, directly applicable UK analysis was available, then other less relevant studies may not have been included. However, in this guideline, no economic studies were excluded on the basis that more applicable evidence was available.

For more details about the assessment of applicability and methodological quality see Table 6 below and the economic evaluation checklist (appendix H of the NICE guidelines manual[4]) and the health economics review protocol, which can be found in each of the evidence reports.

When no relevant health economic studies were found from the economic literature review, relevant UK NHS unit costs related to the compared interventions were presented to the committee to inform the possible economic implications of the recommendations.

#### 2.4.1.2 NICE health economic evidence profiles

NICE health economic evidence profile tables were used to summarise cost and cost-effectiveness estimates for the included health economic studies in each evidence review report. The health economic evidence profile shows an assessment of applicability and methodological quality for each economic study, with footnotes indicating the reasons for the assessment. These assessments were made by the health economist using the economic evaluation checklist from the NICE guidelines manual.[4] It also shows the incremental costs, incremental effects (for example, quality-adjusted life years [QALYs]) and incremental cost-effectiveness ratio (ICER) for the base case analysis in the study, as well as information about the assessment of uncertainty in the analysis. See Table 6 for more details.

When a non-UK study was included in the profile, the results were converted into pounds sterling using the appropriate purchasing power parity.[6]

**Table 6: Content of NICE health economic evidence profile**

| Item | Description |
|---|---|
| Study | Surname of first author, date of study publication and country perspective with a reference to full information on the study. |
| Applicability | An assessment of applicability of the study to this guideline, the current NHS situation and NICE decision-making:[(a)]<br>• Directly applicable – the study meets all applicability criteria, or fails to meet 1 or more applicability criteria but this is unlikely to change the conclusions about cost effectiveness.<br>• Partially applicable – the study fails to meet 1 or more applicability criteria, and this could change the conclusions about cost effectiveness.<br>• Not applicable – the study fails to meet 1 or more of the applicability criteria, and this is likely to change the conclusions about cost |

ISBN: 978-1-4731-3003-6

| Item | Description |
|---|---|
| | effectiveness. Such studies would usually be excluded from the review. |
| Limitations | An assessment of methodological quality of the study:[(a)]<br>• Minor limitations – the study meets all quality criteria, or fails to meet 1 or more quality criteria, but this is unlikely to change the conclusions about cost effectiveness.<br>• Potentially serious limitations – the study fails to meet 1 or more quality criteria, and this could change the conclusions about cost effectiveness.<br>• Very serious limitations – the study fails to meet 1 or more quality criteria, and this is highly likely to change the conclusions about cost effectiveness. Such studies would usually be excluded from the review. |
| Other comments | Information about the design of the study and particular issues that should be considered when interpreting it. |
| Incremental cost | The mean cost associated with one strategy minus the mean cost of a comparator strategy. |
| Incremental effects | The mean QALYs (or other selected measure of health outcome) associated with one strategy minus the mean QALYs of a comparator strategy. |
| Cost effectiveness | Incremental cost-effectiveness ratio (ICER): the incremental cost divided by the incremental effects (usually in £ per QALY gained). |
| Uncertainty | A summary of the extent of uncertainty about the ICER reflecting the results of deterministic or probabilistic sensitivity analyses, or stochastic analyses of trial data, as appropriate. |

*(a) Applicability and limitations were assessed using the economic evaluation checklist in appendix H of the NICE guidelines manual[4]*

### 2.4.2 Cost-effectiveness criteria

NICE's report 'Social value judgements: principles for the development of NICE guidance' sets out the principles that committees should consider when judging whether an intervention offers good value for money.[5] In general, an intervention was considered to be cost effective (given that the estimate was considered plausible) if either of the following criteria applied:

- the intervention dominated other relevant strategies (that is, it was both less costly in terms of resource use and more clinically effective compared with all the other relevant alternative strategies), or
- the intervention cost less than £20,000 per QALY gained compared with the next best strategy.

If the committee recommended an intervention that was estimated to cost more than £20,000 per QALY gained, or did not recommend one that was estimated to cost less than £20,000 per QALY gained, the reasons for this decision are discussed explicitly in 'The committee's discussion of the evidence' section of the relevant evidence report, with reference to issues regarding the plausibility of the estimate or to the factors set out in 'Social value judgements: principles for the development of NICE guidance'.[5]

When QALYs or life years gained are not used in the analysis, results are difficult to interpret unless one strategy dominates the others with respect to every relevant health outcome and cost.

### 2.4.3 In the absence of health economic evidence

When no relevant published health economic studies were found, and a new analysis was not prioritised, the committee made a qualitative judgement about cost effectiveness by considering expected differences in resource use between options and relevant UK NHS unit costs, alongside the results of the review of clinical effectiveness evidence.

ISBN: 978-1-4731-3003-6

The UK NHS costs reported in the guideline are those that were presented to the committee and were correct at the time recommendations were drafted. They may have changed subsequently before the time of publication. However, we have no reason to believe they have changed substantially.

## 2.5 Developing recommendations

Over the course of the guideline development process, the committee was presented with:

- Summaries of clinical and health economic evidence and quality (as presented in evidence reports A–I).
- Evidence tables of the clinical and health economic evidence reviewed from the literature. All evidence tables can be found in appendices to the relevant evidence reports.
- Forest plots (in appendices to the relevant evidence reports).

Recommendations were drafted on the basis of the committee's interpretation of the available evidence, taking into account the balance of benefits, harms and costs between different courses of action. This was either done formally in an economic model, or informally. Firstly, the net clinical benefit over harm (clinical effectiveness) was considered, focusing on the critical outcomes. When this was done informally, the committee took into account the clinical benefits and harms when one intervention was compared with another. The assessment of net clinical benefit was moderated by the importance placed on the outcomes (the committee's values and preferences), and the confidence the committee had in the evidence (evidence quality). Secondly, the committee assessed whether the net clinical benefit justified any differences in costs between the alternative interventions.

When clinical and health economic evidence was of poor quality, conflicting or absent, the committee drafted recommendations based on its expert opinion. The considerations for making consensus-based recommendations include the balance between potential harms and benefits, the economic costs compared to the economic benefits, current practices, recommendations made in other relevant guidelines, patient preferences and equality issues. The consensus recommendations were agreed through discussions in the committee.

All recommendations, those supported by clinical evidence, those agreed through committee consensus and combinations of these approaches, involve uncertainty. The guideline developers assessed the quality of the evidence and the committee discussed their personal uncertainty around their concensus. In doing so the committee also considered whether the uncertainty was sufficient and research lacking to justifymaking a recommendation to forfurther research (see section 2.5.1 below).

The committee considered the appropriate 'strength' of each recommendation. This takes into account the quality of the evidence but is conceptually different. Some recommendations are 'strong' in that the committee believes that the vast majority of healthcare and other professionals and patients would choose a particular intervention if they considered the evidence in the same way that the committee has. This is generally the case if the benefits clearly outweigh the harms for most people and the intervention is likely to be cost effective. However, there is often a closer balance between benefits and harms, and some people would not choose an intervention whereas others would. This may happen, for example, if some patients are particularly averse to some side effect and others are not. In these circumstances the recommendation is generally weaker, although it may be possible to make stronger recommendations about specific groups of people.

The committee focused on the following factors in agreeing the wording of the recommendations:

- The actions health professionals need to take.
- The information readers need to know.

ISBN: 978-1-4731-3003-6

- The strength of the recommendation (for example the word 'offer' was used for strong recommendations and 'consider' for weaker recommendations).
- The involvement of people with rheumatoid arthritis (and their carers if needed) in decisions on treatment and care.
- Consistency with NICE's standard advice on recommendations about drugs, waiting times and ineffective interventions (see section 9.2 in the NICE guidelines manual[4]).

The main considerations specific to each recommendation are outlined in 'The committee's discussion of the evidence' section within each evidence report.

### 2.5.1 Research recommendations

When areas were identified for which good evidence was lacking, the committee considered making recommendations for future research. Decisions about the inclusion of a research recommendation were based on factors such as:

- the importance to patients or the population
- national priorities
- potential impact on the NHS and future NICE guidance
- ethical and technical feasibility.

### 2.5.2 Validation process

This guidance is subject to a 6-week public consultation and feedback as part of the quality assurance and peer review of the document. All comments received from registered stakeholders are responded to in turn and posted on the NICE website.

### 2.5.3 Updating the guideline

Following publication, and in accordance with the NICE guidelines manual, NICE will undertake a review of whether the evidence base has progressed significantly to alter the guideline recommendations and warrant an update.

### 2.5.4 Disclaimer

Healthcare providers need to use clinical judgement, knowledge and expertise when deciding whether it is appropriate to apply guidelines. The recommendations cited here are a guide and may not be appropriate for use in all situations. The decision to adopt any of the recommendations cited here must be made by practitioners in light of individual patient circumstances, the wishes of the patient, clinical expertise and resources.

The National Guideline Centre disclaims any responsibility for damages arising out of the use or non-use of this guideline and the literature used in support of this guideline.

### 2.5.5 Funding

The National Guideline Centre was commissioned by the National Institute for Health and Care Excellence to undertake the work on this guideline.

ISBN: 978-1-4731-3003-6

# 3 Acronyms and abbreviations

| Acronym or abbreviation | Description |
|---|---|
| ACPA | Anti-citrullinated protein antibodies |
| ACR | American College of Rheumatology (see ARA) |
| ACR 20, 50 70 | ACR-criteria 20-50-70 |
| ADL | Activities of daily living |
| AEs | Adverse events |
| AL-TENS | Acupuncture-like transcutaneous electrical nerve stimulation |
| Anti-CCP | Anti-cyclic citrullinated peptide (anti-CCP) antibody |
| Anti-TNF | Anti-tumour necrosis factor |
| ARA | American Rheumatism Association (now ACR) |
| AUC | Area under the curve |
| BMI | Body mass index |
| BSR | British Society of Rheumatology |
| CBT | Cognitive behavioural therapy |
| CCP | Cyclic citrullinated peptide |
| cDMARD | Conventional disease-modifying anti-rheumatic drug |
| CEA | Cost-effectiveness analysis |
| CI | Confidence interval (95% unless stated otherwise) |
| COX-2 | Cyclooxegenase-2 |
| CRP | C-reactive protein |
| CsA | Cyclosporine A |
| CUA | Cost–utility analysis |
| CV | Cardiovascular |
| DAS (DAS28, DAS32) | Disease activity score (disease activity score of 28 or 32 joints, respectively) |
| DMARD | Disease-modifying anti-rheumatic drug |
| DXR BMD | Digital x-ray radiogrammetry bone mineral density |
| EQ-5D/EuroQol | A standardised instrument for use as a measure of health outcome (quality of life) |
| ES | Erosion score |
| ESR | Erythrocyte sedimentation rate |
| EULAR | European League Against Rheumatism |
| GC | Guideline Committee |
| GI | Gastrointestinal |
| GRADE | Grading of recommendations assessment, development and evaluation |
| GS | Grey-scale |
| GSsynSS | Grey-scale synovitis sum score |
| GStenSS | Grey-scale tenosynovitis sum score |
| GSUS | Grey-scale ultrasound |
| HAQ | Health assessment questionnaire |
| HAQ-DI | Health assessment questionnaire disability index |
| Hb | Haemoglobin |
| IA | Intra-articular |
| ICER | Incremental cost-effectiveness ratio |

ISBN: 978-1-4731-3003-6

| Acronym or abbreviation | Description |
|---|---|
| IgA | Immunoglobulin A |
| IgM | Immunoglobulin M |
| IM | Intramuscular |
| IQR | Interquartile range |
| IRGL | Impact of rheumatic disease on general health and lifestyle questionnaire |
| ITT | Intention-to-treat analysis |
| IV | Intravenous |
| LASER | Light amplification by stimulated emission of radiation |
| MACTAR | McMaster Toronto arthritis patient preference disability questionnaire |
| MCP | Metacarpophalangeal joints |
| MDT | Multidisciplinary team |
| mHAQ | Modified Health Assessment Questionnaire |
| MHRA | Medicines and Healthcare Products Regulatory Agency |
| MI | Myocardial infarction |
| MID | Minimal important difference |
| MRI | Magnetic resonance imaging |
| MTP | Metatarsophalangeal joint |
| MVA | Multivariate analysis |
| NCC-CC | National Collaborating Centre for Chronic Conditions (now NGC) |
| NCGC | National Clinical Guideline Centre (now NGC) |
| NGC | National Guideline Centre |
| NICE | National Institute of Health and Care Excellence |
| NS | Not significant (at the 5% level unless stated otherwise) |
| NSAID | Nonsteroidal anti-inflammatory drug |
| OA | Osteoarthritis |
| OECD | Organisation for Economic Co-operation and Development |
| OMERACT | Outcome measures in rheumatology clinical trials |
| OR | Odds ratio |
| OT | Occupational Therapy or Therapist |
| PD | Power Doppler |
| PDA | Power Doppler activity |
| PDsynSS | Power Doppler synovitis sum score |
| PDtenSS | Power Doppler tenosynovitis sum score |
| PDUS | Power Doppler ultrasound |
| PIP | Proximal interphalangeal joints |
| PPI | Proton pump inhibitor |
| PPV | Positive predictive value |
| QALY | Quality-adjusted life year |
| RA | Rheumatoid arthritis |
| RAI | Ritchie Articular Index |
| RAID | Rheumatoid Arthritis Impact of Disease |
| RAMRIS | Rheumatology magnetic resonance imaging scoring system |
| RAQoL | Rheumatoid Arthritis Quality of Life questionnaire |
| RCT | Randomised controlled trial |
| RF | Rheumatoid factor |
| ROM | Range of motion |

ISBN: 978-1-4731-3003-6

| Acronym or abbreviation | Description |
|---|---|
| RR | Relative risk |
| SDAI | Simplified Disease Activity Index |
| SF-12, SF-36 | Short form 12-point or 36-point questionnaire, respectively |
| SJC | Swollen joint count |
| SMD | Standardised mean difference |
| SR | Systematic review |
| SD | Standard deviation |
| SSNRI | Selective serotonin-norepinephrine reuptake inhibitors |
| SSRI | Selective serotonin reuptake inhibitors |
| SvdH | Sharp van der Heijde |
| TENS | Transcutaneous electrical nerve stimulation |
| TJC | Tender joint count |
| TJR | Total joint replacement |
| TSS | Total Sharp score |
| UA | Undifferentiated arthritis |
| UPA | Undifferentiated polyarthritis |
| US | Ultrasound |
| US7 | A semi-quantitative US scoring system combining soft tissue changes (synovitis and tenosynovitis) and erosive bone lesions in seven preselected joints in one US scoring system. |
| UVA | Univariate analysis |
| VAS | Visual analogue scale |
| VASDA | Visual analogue scale disease assessment |
| WMD | Weighted mean differences |

ISBN: 978-1-4731-3003-6

# 4 Glossary

The NICE Glossary can be found at www.nice.org.uk/glossary.

## 4.1 Guideline-specific terms

| Term | Definition |
|---|---|
| ACR (American College of Rheumatology) Criteria: 20, 50 70 | These are criteria to measure the effectiveness of medications or treatments in clinical trials for rheumatoid arthritis. The parameters are: patient assessment, physician assessment, pain scale, disability/functional questionnaire, acute phase reactant (ESR or CRP).<br><br>ACR 20 has a positive outcome if 20% improvement in tender or swollen joint counts were achieved as well as a 20% improvement in at least three of the other five criteria.<br><br>ACR 50 has a positive outcome if 50% improvement in tender or swollen joint counts were achieved as well as a 50% improvement in at least three of the other five criteria.<br><br>ACR 70 has a positive outcome if 70% improvement in tender or swollen joint counts were achieved as well as a 70% improvement in at least three of the other five criteria. |
| Analgesia | Analgesics are sometimes used for relief of pain and stiffness in people with rheumatoid arthritis (RA). Medications considered are non-steroidal anti-inflammatory drugs (NSAIDs), opioids. paracetamol, nefopam, gabapentioniods, tricyclic antidepressants, selective serotonin reuptake inhibitors (SSRI) and serotonin-norepinephrine reuptake inhibitor (SNRI) antidepressants, |
| Anti-TNFa treatment | Tumour necrosis factor alpha or TNFα is a cytokine. Cytokines are substances released by the body during inflammation. Currently, there are five licenced treatments: etanercept, infliximab, adalimumab, certolizumab pegol and golimumab that can block the effect of TNFα. |
| Biological / biologic | Type of DMARD which targets pro-inflammatory cytokines that are involved in joint destruction (particularly TNF-alpha and IL-1). |
| Anti-citrullinated protein antibodies (ACPA) | Autoantibodies that are directed against peptides and proteins that are citrullinated. They are present in the majority of patients with rheumatoid arthritis. |
| Bridging treatment | Glucocorticoids used for a short period of time when a person is starting a new DMARD, intended to improve symtoms while waiting for the new DMARD to take effect (which can take 2 to 3 months). |
| Clinically significant improvement (CSI) | Some trials define a dichotomous outcome of clinically significant pain relief as having been achieved above a specific threshold on a pain score, for example, pain VAS. However, there is no standard threshold and each such trial should be considered individually. |
| C-reative protein (CRP) | An annular, pentameric protein found in blood plasma, whose levels rise in response to inflammation. |
| Anti cyclic citrullinated peptide (anti-CCP) | Anti-cyclic citrullinated peptide (anti-CCP) is an antibody present in most people with rheumatoid arthritis. |
| Disease activity score (DAS) | An assessment used by clinicians to measure rheumatoid arthritis disease activity, to determine whether the signs and symptoms have reduced or stopped. |
| cDMARD | Conventional disease-modifying anti-rheumatic drugs are synthetic drugs that modify disease rather than just alleviating symptoms. They |

ISBN: 978-1-4731-3003-6

| Term | Definition |
|---|---|
| | include methotrexate, sulfasalazine, leflunomide and hydroxychloroquine, but do not include biological DMARDs and targeted synthetic DMARDs. |
| Larsen Score | Method of assessing radiographic joint damage cause by RA. |
| Erythrocyte sedimentation rate (ESR) | The rate at which red blood cells sediment in a period of one hour. |
| Health assessment questionnaire (HAQ) | The Health Assessment Questionnaire (HAQ), published in 1980 by the Stanford Arthritis Center, is a patient reported outcome instrument. It assesses multiple dimensions based on patient-centred values and is a tool for measurement of health status. |
| Health assessment questionnaire disability index (HAQ-DI) | The Health Assessment Questionnaire Disability Index (HAQ-DI) is the disability assessment component of the HAQ, It assesses a person's level of functional ability and includes questions of fine movements of the upper extremity, locomotor activities of the lower extremity, and activities that involve both upper and lower extremities. |
| Low disease activity | Low disease activity was based on a measurement of Disease Activity Score (DAS). Low disease activity, where utilised, was defined by the studies included in the evidence reviews. |
| Palindromic | Palindromi rheumatism is an inflammatory arthritis that causes attacks of joint pain and swelling similar to rheumatoid arthritis. Between attacks the joints return to normal. |
| Parallel combination therapy | Two or more DMARDs commenced at the same time without a step-down strategy. |
| Rapid access to specialist care | Direct access to specialist care without the need of a GP referral. |
| Remission | Remission was either based on Disease Activity Score (DAS) or ACR/EULAR remission criteria. Remission, where utilised, was defined by the studies included in the evidence reviews. |
| Rheumatoid factor (RF) | Rheumatoid factor (RF) is an antibody that is detectable in the blood of approximately 80% of adults with rheumatoid arthritis. |
| Sequential monotherapy | Treatment commencing with a single DMARD that is replaced with a different single DMARD in the case of inadequate response. |
| Step-up strategy | Additional DMARDs are added to DMARD monotherapy when disease is not adequately controlled. |
| Step-down strategy | During treatment with 2 or more DMARDs, tapering and stopping at least 1 drug once disease is adequately controlled. |
| Synovitis | Soft tissue joint swelling. |
| Treat-to-target | A treat-to-target strategy is a strategy that defines a treatment target (such as remission or low disease activity) and applies tight control (for example, monthly visits and respective treatment adjustment) to reach this target. The treatment strategy often follows a protocol for treatment adaptations depending on the disease activity level and degree of response to treatment. |

## 4.2 General terms

| Term | Definition |
|---|---|
| Abstract | Summary of a study, which may be published alone or as an introduction to a full scientific paper. |
| Algorithm (in guidelines) | A flow chart of the clinical decision pathway described in the guideline, where decision points are represented with boxes, linked |

ISBN: 978-1-4731-3003-6

| Term | Definition |
|---|---|
| | with arrows. |
| Allocation concealment | The process used to prevent advance knowledge of group assignment in an RCT. The allocation process should be impervious to any influence by the individual making the allocation, by being administered by someone who is not responsible for recruiting participants. |
| Applicability | How well the results of a study or NICE evidence review can answer a clinical question or be applied to the population being considered. |
| Arm (of a clinical study) | Subsection of individuals within a study who receive one particular intervention, for example placebo arm. |
| Association | Statistical relationship between 2 or more events, characteristics or other variables. The relationship may or may not be causal. |
| Base case analysis | In an economic evaluation, this is the main analysis based on the most plausible estimate of each input. In contrast, see Sensitivity analysis. |
| Baseline | The initial set of measurements at the beginning of a study (after run-in period where applicable), with which subsequent results are compared. |
| Bias | Influences on a study that can make the results look better or worse than they really are. (Bias can even make it look as if a treatment works when it does not.) Bias can occur by chance, deliberately or as a result of systematic errors in the design and execution of a study. It can also occur at different stages in the research process, for example, during the collection, analysis, interpretation, publication or review of research data. For examples see selection bias, performance bias, information bias, confounding factor, and publication bias. |
| Blinding | A way to prevent researchers, doctors and patients in a clinical trial from knowing which study group each patient is in so they cannot influence the results. The best way to do this is by sorting patients into study groups randomly. The purpose of 'blinding' or 'masking' is to protect against bias.<br>A single-blinded study is one in which patients do not know which study group they are in (for example whether they are taking the experimental drug or a placebo). A double-blinded study is one in which neither patients nor the researchers and doctors know which study group the patients are in. A triple blind study is one in which neither the patients, clinicians or the people carrying out the statistical analysis know which treatment patients received. |
| Carer (caregiver) | Someone who looks after family, partners or friends in need of help because they are ill, frail or have a disability. |
| Case–control study | A study to find out the cause(s) of a disease or condition. This is done by comparing a group of patients who have the disease or condition (cases) with a group of people who do not have it (controls) but who are otherwise as similar as possible (in characteristics thought to be unrelated to the causes of the disease or condition). This means the researcher can look for aspects of their lives that differ to see if they may cause the condition.<br>For example, a group of people with lung cancer might be compared with a group of people the same age that do not have lung cancer. The researcher could compare how long both groups had been exposed to tobacco smoke. Such studies are retrospective because they look back in time from the outcome to the possible causes of a disease or condition. |
| Case series | Report of a number of cases of a given disease, usually covering the course of the disease and the response to treatment. There is no |

ISBN: 978-1-4731-3003-6

| Term | Definition |
|---|---|
| | comparison (control) group of patients. |
| Clinical efficacy | The extent to which an intervention is active when studied under controlled research conditions. |
| Clinical effectiveness | How well a specific test or treatment works when used in the 'real world' (for example, when used by a doctor with a patient at home), rather than in a carefully controlled clinical trial. Trials that assess clinical effectiveness are sometimes called management trials.<br>Clinical effectiveness is not the same as efficacy. |
| Clinician | A healthcare professional who provides patient care. For example, a doctor, nurse or physiotherapist. |
| Cochrane Review | The Cochrane Library consists of a regularly updated collection of evidence-based medicine databases including the Cochrane Database of Systematic Reviews (reviews of randomised controlled trials prepared by the Cochrane Collaboration). |
| Cohort study | A study with 2 or more groups of people – cohorts – with similar characteristics. One group receives a treatment, is exposed to a risk factor or has a particular symptom and the other group does not. The study follows their progress over time and records what happens. See also observational study. |
| Comorbidity | A disease or condition that someone has in addition to the health problem being studied or treated. |
| Comparability | Similarity of the groups in characteristics likely to affect the study results (such as health status or age). |
| Concordance | This is a recent term whose meaning has changed. It was initially applied to the consultation process in which doctor and patient agree therapeutic decisions that incorporate their respective views, but now includes patient support in medicine taking as well as prescribing communication. Concordance reflects social values but does not address medicine-taking and may not lead to improved adherence. |
| Confidence interval (CI) | There is always some uncertainty in research. This is because a small group of patients is studied to predict the effects of a treatment on the wider population. The confidence interval is a way of expressing how certain we are about the findings from a study, using statistics. It gives a range of results that is likely to include the 'true' value for the population.<br>The CI is usually stated as '95% CI', which means that the range of values has a 95 in a 100 chance of including the 'true' value. For example, a study may state that "based on our sample findings, we are 95% certain that the 'true' population blood pressure is not higher than 150 and not lower than 110". In such a case the 95% CI would be 110 to 150.<br>A wide confidence interval indicates a lack of certainty about the true effect of the test or treatment – often because a small group of patients has been studied. A narrow confidence interval indicates a more precise estimate (for example, if a large number of patients have been studied). |
| Confounding factor | Something that influences a study and can result in misleading findings if it is not understood or appropriately dealt with.<br>For example, a study of heart disease may look at a group of people that exercises regularly and a group that does not exercise. If the ages of the people in the 2 groups are different, then any difference in heart disease rates between the 2 groups could be because of age rather than exercise. Therefore age is a confounding factor. |
| Consensus methods | Techniques used to reach agreement on a particular issue. Consensus methods may be used to develop NICE guidance if there is not enough good quality research evidence to give a clear answer |

ISBN: 978-1-4731-3003-6

| Term | Definition |
|---|---|
| | to a question. Formal consensus methods include Delphi and nominal group techniques. |
| Control group | A group of people in a study who do not receive the treatment or test being studied. Instead, they may receive the standard treatment (sometimes called 'usual care') or a dummy treatment (placebo). The results for the control group are compared with those for a group receiving the treatment being tested. The aim is to check for any differences.<br>Ideally, the people in the control group should be as similar as possible to those in the treatment group, to make it as easy as possible to detect any effects due to the treatment. |
| Cost–benefit analysis (CBA) | Cost–benefit analysis is one of the tools used to carry out an economic evaluation. The costs and benefits are measured using the same monetary units (for example, pounds sterling) to see whether the benefits exceed the costs. |
| Cost–consequences analysis (CCA) | Cost–consequences analysis is one of the tools used to carry out an economic evaluation. This compares the costs (such as treatment and hospital care) and the consequences (such as health outcomes) of a test or treatment with a suitable alternative. Unlike cost–benefit analysis or cost-effectiveness analysis, it does not attempt to summarise outcomes in a single measure (like the quality-adjusted life year) or in financial terms. Instead, outcomes are shown in their natural units (some of which may be monetary) and it is left to decision-makers to determine whether, overall, the treatment is worth carrying out. |
| Cost-effectiveness analysis (CEA) | Cost-effectiveness analysis is one of the tools used to carry out an economic evaluation. The benefits are expressed in non-monetary terms related to health, such as symptom-free days, heart attacks avoided, deaths avoided or life years gained (that is, the number of years by which life is extended as a result of the intervention). |
| Cost-effectiveness model | An explicit mathematical framework, which is used to represent clinical decision problems and incorporate evidence from a variety of sources in order to estimate the costs and health outcomes. |
| Cost–utility analysis (CUA) | Cost–utility analysis is one of the tools used to carry out an economic evaluation. The benefits are assessed in terms of both quality and duration of life, and expressed as quality-adjusted life years (QALYs). See also utility. |
| Decision analysis | An explicit quantitative approach to decision-making under uncertainty, based on evidence from research. This evidence is translated into probabilities, and then into diagrams or decision trees which direct the clinician through a succession of possible scenarios, actions and outcomes. |
| Deterministic analysis | In economic evaluation, this is an analysis that uses a point estimate for each input. In contrast, see Probabilistic analysis |
| Diagnostic odds ratio | The diagnostic odds ratio is a measure of the effectiveness of a diagnostic test. It is defined as the ratio of the odds of the test being positive if the subject has a disease relative to the odds of the test being positive if the subject does not have the disease. |
| Discounting | Costs and perhaps benefits incurred today have a higher value than costs and benefits occurring in the future. Discounting health benefits reflects individual preference for benefits to be experienced in the present rather than the future. Discounting costs reflects individual preference for costs to be experienced in the future rather than the present. |
| Disutility | The loss of quality of life associated with having a disease or condition. See Utility |

ISBN: 978-1-4731-3003-6

| Term | Definition |
|---|---|
| Dominance | A health economics term. When comparing tests or treatments, an option that is both less effective and costs more is said to be 'dominated' by the alternative. |
| Drop-out | A participant who withdraws from a trial before the end. |
| Economic evaluation | An economic evaluation is used to assess the cost effectiveness of healthcare interventions (that is, to compare the costs and benefits of a healthcare intervention to assess whether it is worth doing). The aim of an economic evaluation is to maximise the level of benefits – health effects – relative to the resources available. It should be used to inform and support the decision-making process; it is not supposed to replace the judgement of healthcare professionals.<br>There are several types of economic evaluation: cost–benefit analysis, cost–consequences analysis, cost-effectiveness analysis, cost-minimisation analysis and cost–utility analysis. They use similar methods to define and evaluate costs, but differ in the way they estimate the benefits of a particular drug, programme or intervention. |
| Effect<br>(as in effect measure, treatment effect, estimate of effect, effect size) | A measure that shows the magnitude of the outcome in one group compared with that in a control group.<br>For example, if the absolute risk reduction is shown to be 5% and it is the outcome of interest, the effect size is 5%.<br>The effect size is usually tested, using statistics, to find out how likely it is that the effect is a result of the treatment and has not just happened by chance (that is, to see if it is statistically significant). |
| Effectiveness | How beneficial a test or treatment is under usual or everyday conditions, compared with doing nothing or opting for another type of care. |
| Efficacy | How beneficial a test, treatment or public health intervention is under ideal conditions (for example, in a laboratory), compared with doing nothing or opting for another type of care. |
| Epidemiological study | The study of a disease within a population, defining its incidence and prevalence and examining the roles of external influences (for example, infection, diet) and interventions. |
| EQ-5D (EuroQol 5 dimensions) | A standardised instrument used to measure health-related quality of life. It provides a single index value for health status. |
| Evidence | Information on which a decision or guidance is based. Evidence is obtained from a range of sources including randomised controlled trials, observational studies, expert opinion (of clinical professionals or patients). |
| Exclusion criteria (literature review) | Explicit standards used to decide which studies should be excluded from consideration as potential sources of evidence. |
| Exclusion criteria (clinical study) | Criteria that define who is not eligible to participate in a clinical study. |
| Extrapolation | An assumption that the results of studies of a specific population will also hold true for another population with similar characteristics. |
| Follow-up | Observation over a period of time of an individual, group or initially defined population whose appropriate characteristics have been assessed in order to observe changes in health status or health-related variables. |
| Generalisability | The extent to which the results of a study hold true for groups that did not participate in the research. See also external validity. |
| Gold standard | A method, procedure or measurement that is widely accepted as being the best available to test for or treat a disease. |
| GRADE, GRADE profile | A system developed by the GRADE Working Group to address the shortcomings of present grading systems in healthcare. The GRADE system uses a common, sensible and transparent approach to |

ISBN: 978-1-4731-3003-6

| Term | Definition |
|---|---|
| | grading the quality of evidence. The results of applying the GRADE system to clinical trial data are displayed in a table known as a GRADE profile. |
| Harms | Adverse effects of an intervention. |
| Health economics | Study or analysis of the cost of using and distributing healthcare resources. |
| Health-related quality of life (HRQoL) | A measure of the effects of an illness to see how it affects someone's day-to-day life. |
| Heterogeneity<br>or Lack of homogeneity | The term is used in meta-analyses and systematic reviews to describe when the results of a test or treatment (or estimates of its effect) differ significantly in different studies. Such differences may occur as a result of differences in the populations studied, the outcome measures used or because of different definitions of the variables involved. It is the opposite of homogeneity. |
| Imprecision | Results are imprecise when studies include relatively few patients and few events and thus have wide confidence intervals around the estimate of effect. |
| Inclusion criteria (literature review) | Explicit criteria used to decide which studies should be considered as potential sources of evidence. |
| Incremental cost-effectiveness ratio (ICER) | The difference in the mean costs in the population of interest divided by the differences in the mean outcomes in the population of interest for one treatment compared with another. |
| Indirectness | The available evidence is different to the review question being addressed, in terms of PICO (population, intervention, comparison and outcome). |
| Intention-to-treat analysis (ITT) | An assessment of the people taking part in a clinical trial, based on the group they were initially (and randomly) allocated to. This is regardless of whether or not they dropped out, fully complied with the treatment or switched to an alternative treatment. Intention-to-treat analyses are often used to assess clinical effectiveness because they mirror actual practice: that is, not everyone complies with treatment and the treatment people receive may be changed according to how they respond to it. |
| Intervention | In medical terms this could be a drug treatment, surgical procedure, diagnostic or psychological therapy. Examples of public health interventions could include action to help someone to be physically active or to eat a more healthy diet. |
| Licence | See 'Product licence'. |
| Life years gained | Mean average years of life gained per person as a result of the intervention compared with an alternative intervention. |
| Likelihood ratio | The likelihood ratio combines information about the sensitivity and specificity. It tells you how much a positive or negative result changes the likelihood that a patient would have the disease. The likelihood ratio of a positive test result (LR+) is sensitivity divided by (1 minus specificity). |
| Loss to follow-up | A patient, or the proportion of patients, actively participating in a clinical trial at the beginning, but whom the researchers were unable to trace or contact by the point of follow-up in the trial |
| Markov model | A method for estimating long-term costs and effects for recurrent or chronic conditions, based on health states and the probability of transition between them within a given time period (cycle). |
| Meta-analysis | A method often used in systematic reviews. Results from several studies of the same test or treatment are combined to estimate the overall effect of the treatment. |
| Multivariate model | A statistical model for analysis of the relationship between 2 or more |

ISBN: 978-1-4731-3003-6

| Term | Definition |
|---|---|
| | predictor (independent) variables and the outcome (dependent) variable. |
| Negative predictive value (NPV) | In screening or diagnostic tests: A measure of the usefulness of a screening or diagnostic test. It is the proportion of those with a negative test result who do not have the disease, and can be interpreted as the probability that a negative test result is correct. It is calculated as follows: TN/(TN+FN) |
| Non-randomised intervention study | A quantitative study investigating the effectiveness of an intervention that does not use randomisation to allocate patients (or units) to treatment groups. Non-randomised studies include observational studies, where allocation to groups occurs through usual treatment decisions or people's preferences. Non-randomised studies can also be experimental, where the investigator has some degree of control over the allocation of treatments.<br>Non-randomised intervention studies can use a number of different study designs, and include cohort studies, case–control studies, controlled before-and-after studies, interrupted-time-series studies and quasi-randomised controlled trials. |
| Number needed to treat (NNT) | The average number of patients who need to be treated to get a positive outcome. For example, if the NNT is 4, then 4 patients would have to be treated to ensure 1 of them gets better. The closer the NNT is to 1, the better the treatment.<br>For example, if you give a stroke prevention drug to 20 people before 1 stroke is prevented, the number needed to treat is 20. See also number needed to harm, absolute risk reduction. |
| Observational study | Individuals or groups are observed or certain factors are measured. No attempt is made to affect the outcome. For example, an observational study of a disease or treatment would allow 'nature' or usual medical care to take its course. Changes or differences in one characteristic (for example, whether or not people received a specific treatment or intervention) are studied without intervening.<br>There is a greater risk of selection bias than in experimental studies. |
| Odds ratio | Odds are a way to represent how likely it is that something will happen (the probability). An odds ratio compares the probability of something in one group with the probability of the same thing in another.<br>An odds ratio of 1 between 2 groups would show that the probability of the event (for example a person developing a disease, or a treatment working) is the same for both. An odds ratio greater than 1 means the event is more likely in the first group. An odds ratio less than 1 means that the event is less likely in the first group.<br>Sometimes probability can be compared across more than 2 groups – in this case, one of the groups is chosen as the 'reference category', and the odds ratio is calculated for each group compared with the reference category. For example, to compare the risk of dying from lung cancer for non-smokers, occasional smokers and regular smokers, non-smokers could be used as the reference category. Odds ratios would be worked out for occasional smokers compared with non-smokers and for regular smokers compared with non-smokers. See also confidence interval, risk ratio. |
| Opportunity cost | The loss of other healthcare programmes displaced by investment in or introduction of another intervention. This may be best measured by the health benefits that could have been achieved had the money been spent on the next best alternative healthcare intervention. |
| Outcome | The impact that a test, treatment, policy, programme or other intervention has on a person, group or population. Outcomes from interventions to improve the public's health could include changes in |

ISBN: 978-1-4731-3003-6

| Term | Definition |
|---|---|
| | knowledge and behaviour related to health, societal changes (for example, a reduction in crime rates) and a change in people's health and wellbeing or health status. In clinical terms, outcomes could include the number of patients who fully recover from an illness or the number of hospital admissions, and an improvement or deterioration in someone's health, functional ability, symptoms or situation. Researchers should decide what outcomes to measure before a study begins. |
| p value | The p value is a statistical measure that indicates whether or not an effect is statistically significant.<br>For example, if a study comparing 2 treatments found that one seems more effective than the other, the p value is the probability of obtaining these results by chance. By convention, if the p value is below 0.05 (that is, there is less than a 5% probability that the results occurred by chance) it is considered that there probably is a real difference between treatments. If the p value is 0.001 or less (less than a 1% probability that the results occurred by chance), the result is seen as highly significant.<br>If the p value shows that there is likely to be a difference between treatments, the confidence interval describes how big the difference in effect might be. |
| Placebo | A fake (or dummy) treatment given to participants in the control group of a clinical trial. It is indistinguishable from the actual treatment (which is given to participants in the experimental group). The aim is to determine what effect the experimental treatment has had – over and above any placebo effect caused because someone has received (or thinks they have received) care or attention. |
| Polypharmacy | The use or prescription of multiple medications. |
| Positive predictive value (PPV) | In screening or diagnostic tests: A measure of the usefulness of a screening or diagnostic test. It is the proportion of those with a positive test result who have the disease, and can be interpreted as the probability that a positive test result is correct. It is calculated as follows: TP/(TP+FP) |
| Post-test probability | In diagnostic tests: The proportion of patients with that particular test result who have the target disorder (post-test odds/[1 plus post-test odds]). |
| Power (statistical) | The ability to demonstrate an association when one exists. Power is related to sample size; the larger the sample size, the greater the power and the lower the risk that a possible association could be missed. |
| Pre-test probability | In diagnostic tests: The proportion of people with the target disorder in the population at risk at a specific time point or time interval. Prevalence may depend on how a disorder is diagnosed. |
| Prevalence | See Pre-test probability. |
| Primary care | Healthcare delivered outside hospitals. It includes a range of services provided by GPs, nurses, health visitors, midwives and other healthcare professionals and allied health professionals such as dentists, pharmacists and opticians. |
| Primary outcome | The outcome of greatest importance, usually the one in a study that the power calculation is based on. |
| Probabilistic analysis | In economic evaluation, this is an analysis that uses a probability distribution for each input. In contrast, see Deterministic analysis. |
| Product licence | An authorisation from the MHRA to market a medicinal product. |
| Prognosis | A probable course or outcome of a disease. Prognostic factors are patient or disease characteristics that influence the course. Good prognosis is associated with low rate of undesirable outcomes; poor |

ISBN: 978-1-4731-3003-6

| Term | Definition |
|---|---|
| | prognosis is associated with a high rate of undesirable outcomes. |
| Prospective study | A research study in which the health or other characteristic of participants is monitored (or 'followed up') for a period of time, with events recorded as they happen. This contrasts with retrospective studies. |
| Publication bias | Publication bias occurs when researchers publish the results of studies showing that a treatment works well and don't publish those showing it did not have any effect. If this happens, analysis of the published results will not give an accurate idea of how well the treatment works. This type of bias can be assessed by a funnel plot. |
| Quality of life | See 'Health-related quality of life'. |
| Quality-adjusted life year (QALY) | A measure of the state of health of a person or group in which the benefits, in terms of length of life, are adjusted to reflect the quality of life. One QALY is equal to 1 year of life in perfect health.<br>QALYS are calculated by estimating the years of life remaining for a patient following a particular treatment or intervention and weighting each year with a quality of life score (on a scale of 0 to 1). It is often measured in terms of the person's ability to perform the activities of daily life, freedom from pain and mental disturbance. |
| Randomisation | Assigning participants in a research study to different groups without taking any similarities or differences between them into account. For example, it could involve using a random numbers table or a computer-generated random sequence. It means that each individual (or each group in the case of cluster randomisation) has the same chance of receiving each intervention. |
| Randomised controlled trial (RCT) | A study in which a number of similar people are randomly assigned to 2 (or more) groups to test a specific drug or treatment. One group (the experimental group) receives the treatment being tested, the other (the comparison or control group) receives an alternative treatment, a dummy treatment (placebo) or no treatment at all. The groups are followed up to see how effective the experimental treatment was. Outcomes are measured at specific times and any difference in response between the groups is assessed statistically. This method is also used to reduce bias. |
| Receiver operated characteristic (ROC) curve | A graphical method of assessing the accuracy of a diagnostic test. Sensitivity is plotted against 1 minus specificity. A perfect test will have a positive, vertical linear slope starting at the origin. A good test will be somewhere close to this ideal. |
| Reference standard | The test that is considered to be the best available method to establish the presence or absence of the outcome – this may not be the one that is routinely used in practice. |
| Reporting bias | See 'Publication bias'. |
| Resource implication | The likely impact in terms of finance, workforce or other NHS resources. |
| Retrospective study | A research study that focuses on the past and present. The study examines past exposure to suspected risk factors for the disease or condition. Unlike prospective studies, it does not cover events that occur after the study group is selected. |
| Review question | In guideline development, this term refers to the questions about treatment and care that are formulated to guide the development of evidence-based recommendations. |
| Risk ratio (RR) | The ratio of the risk of disease or death among those exposed to certain conditions compared with the risk for those who are not exposed to the same conditions (for example, the risk of people who smoke getting lung cancer compared with the risk for people who do not smoke). |

ISBN: 978-1-4731-3003-6

| Term | Definition |
|---|---|
| | If both groups face the same level of risk, the risk ratio is 1. If the first group had a risk ratio of 2, subjects in that group would be twice as likely to have the event happen. A risk ratio of less than 1 means the outcome is less likely in the first group. The risk ratio is sometimes referred to as relative risk. |
| Secondary outcome | An outcome used to evaluate additional effects of the intervention deemed a priori as being less important than the primary outcomes. |
| Selection bias | Selection bias occurs if:<br>a) The characteristics of the people selected for a study differ from the wider population from which they have been drawn, or<br>b) There are differences between groups of participants in a study in terms of how likely they are to get better. |
| Sensitivity | How well a test detects the thing it is testing for.<br>If a diagnostic test for a disease has high sensitivity, it is likely to pick up all cases of the disease in people who have it (that is, give a 'true positive' result). But if a test is too sensitive it will sometimes also give a positive result in people who don't have the disease (that is, give a 'false positive').<br>For example, if a test were developed to detect if a woman is 6 months pregnant, a very sensitive test would detect everyone who was 6 months pregnant, but would probably also include those who are 5 and 7 months pregnant.<br>If the same test were more specific (sometimes referred to as having higher specificity), it would detect only those who are 6 months pregnant, and someone who was 5 months pregnant would get a negative result (a 'true negative'). But it would probably also miss some people who were 6 months pregnant (that is, give a 'false negative').<br>Breast screening is a 'real-life' example. The number of women who are recalled for a second breast screening test is relatively high because the test is very sensitive. If it were made more specific, people who don't have the disease would be less likely to be called back for a second test but more women who have the disease would be missed. |
| Sensitivity analysis | A means of representing uncertainty in the results of economic evaluations. Uncertainty may arise from missing data, imprecise estimates or methodological controversy. Sensitivity analysis also allows for exploring the generalisability of results to other settings. The analysis is repeated using different assumptions to examine the effect on the results.<br>One-way simple sensitivity analysis (univariate analysis): each parameter is varied individually in order to isolate the consequences of each parameter on the results of the study.<br>Multi-way simple sensitivity analysis (scenario analysis): 2 or more parameters are varied at the same time and the overall effect on the results is evaluated.<br>Threshold sensitivity analysis: the critical value of parameters above or below which the conclusions of the study will change are identified.<br>Probabilistic sensitivity analysis: probability distributions are assigned to the uncertain parameters and are incorporated into evaluation models based on decision analytical techniques (for example, Monte Carlo simulation). |
| Significance (statistical) | A result is deemed statistically significant if the probability of the result occurring by chance is less than 1 in 20 ($p<0.05$). |
| Specificity | The proportion of true negatives that are correctly identified as such. For example in diagnostic testing the specificity is the proportion of |

ISBN: 978-1-4731-3003-6

| Term | Definition |
|---|---|
| | non-cases correctly diagnosed as non-cases.<br>See related term ‘Sensitivity’.<br>In terms of literature searching a highly specific search is generally narrow and aimed at picking up the key papers in a field and avoiding a wide range of papers. |
| Stakeholder | An organisation with an interest in a topic that NICE is developing a guideline or piece of public health guidance on. Organisations that register as stakeholders can comment on the draft scope and the draft guidance. Stakeholders may be:<br>• manufacturers of drugs or equipment<br>• national patient and carer organisations<br>• NHS organisations<br>• organisations representing healthcare professionals. |
| Systematic review | A review in which evidence from scientific studies has been identified, appraised and synthesised in a methodical way according to predetermined criteria. It may include a meta-analysis. |
| Technology appraisal | Formal ascertainment and review of the evidence surrounding a health technology, restricted in the current document to appraisals undertaken by NICE. |
| Time horizon | The time span over which costs and health outcomes are considered in a decision analysis or economic evaluation. |
| Transition probability | In a state transition model (Markov model), this is the probability of moving from one health state to another over a specific period of time. |
| Treatment allocation | Assigning a participant to a particular arm of a trial. |
| Univariate | Analysis which separately explores each variable in a data set. |
| Utility | In health economics, a 'utility' is the measure of the preference or value that an individual or society places upon a particular health state. It is generally a number between 0 (representing death) and 1 (perfect health). The most widely used measure of benefit in cost–utility analysis is the quality-adjusted life year, but other measures include disability-adjusted life years (DALYs) and healthy year equivalents (HYEs). |

ISBN: 978-1-4731-3003-6

# References

1. GRADE Working Group. The Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group website. 2011. Available from: http://www.gradeworkinggroup.org/ Last accessed:

2. Hayden JA, van der Windt DA, Cartwright JL, Cote P, Bombardier C. Assessing bias in studies of prognostic factors. Annals of Internal Medicine. 2013; 158(4):280-286

3. National Collaborating Centre for Chronic Conditions. Rheumatoid arthritis: national clinical guideline for management and treatment in adults. NICE clinical guideline 79. London. Royal College of Physicians, 2009. Available from: http://guidance.nice.org.uk/CG79

4. National Institute for Health and Care Excellence. Developing NICE guidelines: the manual. London. National Institute for Health and Care Excellence, 2014. Available from: http://www.nice.org.uk/article/PMG20/chapter/1%20Introduction%20and%20overview

5. National Institute for Health and Clinical Excellence. Social value judgements: principles for the development of NICE guidance. London. National Institute for Health and Clinical Excellence, 2008. Available from: https://www.nice.org.uk/media/default/about/what-we-do/research-and-development/social-value-judgements-principles-for-the-development-of-nice-guidance.pdf

6. Organisation for Economic Co-operation and Development (OECD). Purchasing power parities (PPP). 2015. Available from: http://www.oecd.org/std/prices-ppp/ Last accessed: 15/11/2017.

7. Review Manager (RevMan) [Computer program]. Version 5. Copenhagen. The Nordic Cochrane Centre, The Cochrane Collaboration, 2015. Available from: http://tech.cochrane.org/Revman

ISBN: 978-1-4731-3003-6