National Institute for Health and Care Excellence

Guideline version (Final)

Gout: diagnosis and management

NICE guideline: Methodology

NICE guideline NG219

Methods

June 2022

Final

National Institute for Health and Care Excellence

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 All NICE guidance is subject to regular review and may be updated or withdrawn.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: 978-1-4731-4603-7

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:

- management of gout

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 Dr Aung Soe 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.

The committee invited Dr Samuel Finnikin, National Clinical Specialist Advisor in Personalised Care, GP for Sutton Coldfield Group Practice and Clinical Research Fellow at the University of Birmingham, to peer review the draft guideline and comment during the stakeholder consultation between the 13/12/21 and 2/2/22.

1.3.1 What this guideline covers

The diagnosis and management of gout in adults 18 years and older.

The key areas covered are:

- information and support for people with gout and their families or carers
- diagnosis and assessment of gout
- pharmacological and non-pharmacological management of gout flares
- long-term management of gout including urate lowering therapy and diet and lifestyle
- ongoing care and monitoring
- referral to specialist services.

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

People with calcium pyrophosphate crystal deposition, including pseudogout.

Relationships between the guideline and other NICE guidance

NICE technology appraisals to be updated by this guidance:

- Febuxostat for the management of hyperuricaemia in people with gout. NICE technology appraisal guidance 164 (2008).
- Canakinumab for treating gouty arthritis attacks and reducing the frequency of subsequent attacks (terminated appraisal 281 (2013)

Related NICE technology appraisals:

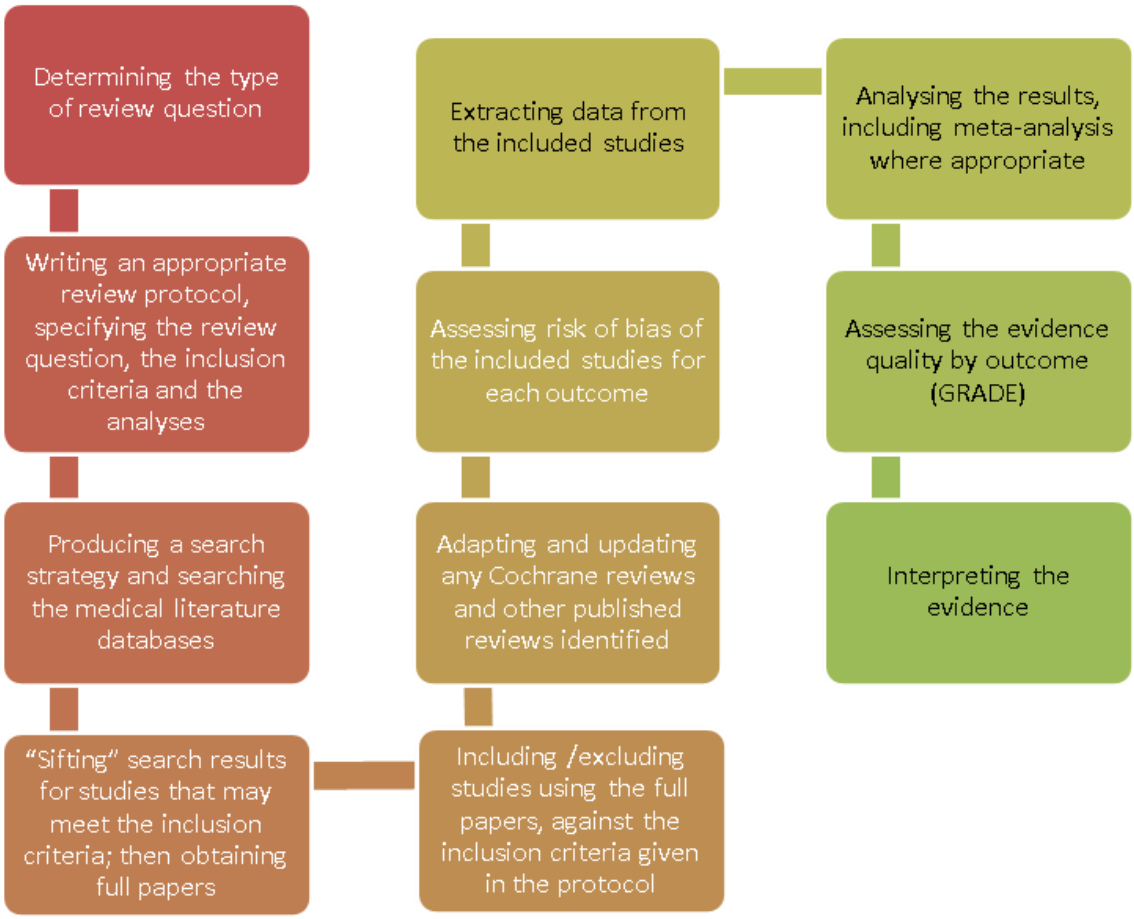
- Lesinurad for treating chronic hyperuricaemia in people with gout. NICE technology appraisal guidance 506 (2018).

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, updated 2020.[10]

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; and using a framework of population, setting and context for qualitative 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.

A total of 15 review questions were identified.

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]1.1	Qualitative	What information and support is needed by people with gout and their families or carers in relation to gout, and when should this be provided?	Themes will be derived from the evidence identified for this review and not pre-specified but areas of interest include: Who will provide information? (preference) Format (written, internet) Delivery (face-to-face, telephone, video conferencing, one-one one versus group) How often? And when delivered?
[B]2.1	Diagnostic	What signs and symptoms indicate gout as a possible diagnosis?	Diagnostic accuracy review: Primary paired outcome: • Sensitivity/specificity If no diagnostic accuracy studies are found, we will look for diagnostic association studies: • Association data: ○ Adjusted RR or OR (adjusted for key confounders of age or gender)
[C]2.2	Diagnostic	What are the most accurate and cost-effective approaches to diagnosing gout, in particular serum urate level compared with joint aspiration?	Primary paired outcome: Sensitivity/specificity
[D]3.1-3.2	Intervention	What is the clinical and cost effectiveness of pharmacological interventions (including NSAIDs, colchicine, corticosteroids and IL-1 inhibitors) and non-	All outcomes are considered equally important for decision making and therefore have all been rated as critical:

Evidence report	Type of review	Review questions	Outcomes
		pharmacological interventions for managing gout flares?	• health-related quality of life (e.g. as described by SF-36, Gout Assessment Questionnaire (GAQ) and the Gout Impact Scale (GIS) or other validated gout-specific HRQoL measures • pain (measured on a visual analogue scale (VAS) or numerical rating scale such as the five-point Likert scale, or reported as pain relief of 50% or greater) • joint swelling/ joint inflammation • joint tenderness • patient global assessment of treatment success (response to treatment) (e.g. Likert scales, visual analogue scales (VAS), numerical ratings scales (NRS)) • adverse events – (1) cardiovascular, (2) renal and (3) gastrointestinal (e.g. diarrhoea) (total adverse events will be reported if the specific types of adverse events are not reported) • admissions (hospital and A&E) • GP visits Timepoints: short (up to two weeks), medium (two to six weeks) and long (> six weeks) term
[E]4.1a	Prognostic	Which people with gout should be offered a urate-lowering therapy such as a xanthine oxidase inhibitor, a uricosuric or a uricase?	All outcomes are considered equally important for decision making and therefore have all been rated as critical: • Frequency of flares • Health-related quality of life (e.g. as described by SF-36,

Evidence report	Type of review	Review questions	Outcomes
			Gout Assessment Questionnaire (GAQ) and the Gout Impact Scale (GIS) or other validated gout-specific HRQoL measures
[F]4.1b	Intervention	When should urate-lowering therapy be started, in relation to a flare, in people with gout?	All outcomes are considered equally important for decision making and therefore have all been rated as critical: • health-related quality of life (e.g. as described by SF-36, Gout Assessment Questionnaire (GAQ) and the Gout Impact Scale (GIS) or other validated gout-specific HRQoL measures • pain (measured on a visual analogue scale (VAS) or numerical rating scale such as the five-point Likert scale, or reported as pain relief of 50% or greater) • joint swelling/joint inflammation • joint tenderness • frequency of flares • flare duration • patient global assessment of treatment success (response to treatment) (e.g. Likert scales, visual analogue scales (VAS), numerical ratings scales (NRS)) • serum urate levels • admissions (hospital and A&E/urgent care) • GP visits Timepoints: short (up to two weeks), medium (two to six weeks) and long (> six weeks) term

Evidence report	Type of review	Review questions	Outcomes

[G]4.2
 	
This protocol was merged with protocol 4.3-4.4 into one review
	Intervention	In people with gout (including people with gout and chronic kidney disease), which urate-lowering therapies (either alone or in combination with each other) are the most clinically and cost effective for first-line treatment?	All outcomes are considered equally important for decision making and therefore have all been rated as critical:
			• health-related quality of life (e.g. as described by SF-36, Gout Assessment Questionnaire (GAQ) and the Gout Impact Scale (GIS) or other validated gout-specific HRQoL measures
			• pain (measured on a visual analogue scale (VAS) or numerical rating scale such as the five-point Likert scale, or reported as pain relief of 50% or greater)
			• joint swelling/joint inflammation
			• joint tenderness
			• frequency of flares
			• patient global assessment of treatment success (response to treatment) (e.g. Likert scales, visual analogue scales (VAS), numerical ratings scales (NRS))
			• adverse events – (1) cardiovascular, (2) renal and (3) gastrointestinal (e.g. diarrhoea) (Total adverse events will be reported if the specific types of adverse events are not reported)
			• adverse events and complications of gout:
			  ○ radiographic joint damage
			  ○ renal stones
			  ○ tophi
			• serum urate levels

Evidence report	Type of review	Review questions	Outcomes
			• admissions (hospital and A&E) • GP visits Timepoints: short-term (less than three months), medium-term (three to 12 months) and long-term (more than 12 months) duration
[G]4.3-4.4 This protocol was merged with protocol 4.2 into one review	Intervention	In people with gout (including people with gout and chronic kidney disease), which urate-lowering therapies (either alone or in combination with each other) are the most clinically and cost effective as second line treatment if first line is not tolerated or provides inadequate control?	All outcomes are considered equally important for decision making and therefore have all been rated as critical: • health-related quality of life (e.g. as described by SF-36, Gout Assessment Questionnaire (GAQ) and the Gout Impact Scale (GIS) or other validated gout-specific HRQoL measures • pain (measured on a visual analogue scale (VAS) or numerical rating scale such as the five-point Likert scale, or reported as pain relief of 50% or greater) • joint swelling/joint inflammation • joint tenderness • frequency of flares • patient global assessment of treatment success (response to treatment) (e.g. Likert scales, visual analogue scales (VAS), numerical ratings scales (NRS)) • adverse events – (1) cardiovascular, (2) renal and (3) gastrointestinal (e.g. diarrhoea) (total adverse events will be reported if the specific types of adverse events are not reported)

Evidence report	Type of review	Review questions	Outcomes

			• adverse events and complications of gout:
			  ○ radiographic joint damage
			  ○ renal stones
			  ○ tophi
			• serum urate levels
			• admissions (hospital and A&E/urgent care)
			• GP visits
			Timepoints: short-term (less than three months), medium-term (three to 12 months) and long-term (more than 12 months) duration.

[H]4.5	Intervention	In people with gout (including people with gout and chronic kidney disease), what is the clinical and cost effectiveness of colchicine, NSAIDs, corticosteroids and IL-1 inhibitors for the prevention of gout flares during the initiation or titration of urate-lowering therapy?	All outcomes are considered equally important for decision making and therefore have all been rated as critical: • health-related quality of life (e.g. as described by SF-36, Gout Assessment Questionnaire (GAQ) and the Gout Impact Scale (GIS) or other validated gout-specific HRQoL measures • pain (measured on a visual analogue scale (VAS) or numerical rating scale such as the five-point Likert scale, or reported as pain relief of 50% or greater) • joint swelling/joint inflammation • joint tenderness • frequency of flares • patient global assessment of treatment success (response to treatment) (e.g. Likert scales, visual analogue scales (VAS), numerical ratings scales (NRS)) • serum urate levels

Evidence report	Type of review	Review questions	Outcomes
			• adverse events – (1) cardiovascular, (2) renal and (3) gastrointestinal (e.g. diarrhoea) (Total adverse events will be reported if the specific types of adverse events are not reported) • admission (hospital and A&E) • discontinuation of ULT Timepoints: short (up to two weeks), medium (two to six weeks) and long (> six weeks) term
[I]4.6	Intervention	What is the clinical and cost effectiveness of diet and lifestyle modifications for gout?	All outcomes are considered equally important for decision making and therefore have all been rated as critical: • health-related quality of life (e.g. as described by SF-36, Gout Assessment Questionnaire (GAQ) and the Gout Impact Scale (GIS) or other validated gout-specific HRQoL measures (we would prioritise the health-related quality of life measures listed, but if they are not reported we would look at other validated gout-specific HRQoL measures). • pain (measured on a visual analogue scale (VAS) or numerical rating scale such as the five-point Likert scale, or reported as pain relief of 50% or greater) • joint swelling/joint inflammation • joint tenderness • frequency of flares • patient global assessment of

Evidence report	Type of review	Review questions	Outcomes

			treatment success (response to treatment) (e.g. Likert scales, visual analogue scales (VAS), numerical ratings scales (NRS))
			• adverse events and complications of gout:
			  ○ radiographic joint damage
			  ○ renal stones
			  ○ tophi
			(total adverse events will be reported if the specific adverse events are not reported)
			• serum urate levels
			• admissions (hospital and A&E/urgent care)
			• GP visits
			
			Timepoints: short-term (less than three months), medium-term (three to 12 months) and long-term (more than 12 months) duration

[J]5.1	Intervention	What is the clinical and cost effectiveness of a 'treat-to-target' urate lowering management strategy compared with usual care for gout?	All outcomes are considered equally important for decision making and therefore have all been rated as critical: • health-related quality of life (e.g. as described by SF-36, Gout Assessment Questionnaire (GAQ) and the Gout Impact Scale (GIS) or other validated gout-specific HRQoL measures • pain (measured on a visual analogue scale (VAS) or numerical rating scale such as the five-point Likert scale, or reported as pain relief of 50% or greater) • joint swelling/joint inflammation • joint tenderness • frequency of flares

Evidence report	Type of review	Review questions	Outcomes

			• patient global assessment of treatment success (response to treatment) (e.g. Likert scales, visual analogue scales (VAS), numerical ratings scales (NRS))
			• adverse events – (1) cardiovascular, (2) renal and (3) gastrointestinal (e.g. diarrhoea) (total adverse events will be reported if the specific types of adverse events are not reported) (cardiovascular events can include stroke and coronary artery disease)
			• adverse events and complications of gout:
			  ○ radiographic joint damage
			  ○ renal stones
			  ○ tophi
			• admissions (hospital and A&E/urgent care)
			• GP visits
			
			Timepoints: short-term (less than three months), medium-term (three to 12 months) and long-term (more than 12 months) duration.

[K]5.2	Intervention	What is the best serum urate level target to use when treating-to-target in gout?	All outcomes are considered equally important for decision making and therefore have all been rated as critical: • health-related quality of life (e.g. as described by SF-36, Gout Assessment Questionnaire (GAQ) and the Gout Impact Scale (GIS) or other validated gout-specific HRQoL measures

Evidence report	Type of review	Review questions	Outcomes
			• patient global assessment of treatment success (response to treatment) (e.g. Likert scales, visual analogue scales (VAS), numerical ratings scales (NRS)) • pain (measured on a visual analogue scale (VAS) or numerical rating scale such as the five-point Likert scale, or reported as pain relief of 50% or greater) • joint swelling/joint inflammation • joint tenderness • proportion of participants who reach serum urate target level • frequency of flares • tophi • admissions (hospital and A&E/urgent care) • GP visits Timepoints: short-term (less than three months), medium-term (three to 12 months) and long-term (more than 12 months) duration.
[L]5.3	Intervention	What is the optimum frequency of serum urate level monitoring for people continuing on urate-lowering therapies for gout?	All outcomes are considered equally important for decision making and therefore have all been rated as critical: • health-related quality of life (e.g. as described by SF-36, Gout Assessment Questionnaire (GAQ) and the Gout Impact Scale (GIS) or other validated gout-specific HRQoL measures • patient global assessment of treatment success (response to treatment) (e.g. Likert

Evidence report	Type of review	Review questions	Outcomes
			scales, visual analogue scales (VAS), numerical ratings scales (NRS)) • pain (measured on a visual analogue scale (VAS) or numerical rating scale such as the five-point Likert scale, or reported as pain relief of 50% or greater) • joint swelling/joint inflammation • joint tenderness • serum urate level • frequency of flares • tophi • admissions (hospital and A&E/urgent care) • GP visits Timepoints: Short-term 6 months, medium 6-12 months, long-term 12+ months
[M]5.4	Intervention	What follow-up should be offered to people with gout after a gout flare?	All outcomes are considered equally important for decision making and therefore have all been rated as critical: • health-related quality of life (e.g. as described by SF-36, Gout Assessment Questionnaire (GAQ) and the Gout Impact Scale (GIS) or other validated gout-specific HRQoL measures • frequency of flares • patient global assessment of treatment success (response to treatment) (e.g. Likert scales, visual analogue scales (VAS), numerical ratings scales (NRS)) • proportion of people with gout using ULT

Evidence report	Type of review	Review questions	Outcomes
			• patient awareness of their condition/treatment • serum urate levels • admissions (hospital and A&E/urgent care) • GP visits Timepoints: short-term (less than three months), medium-term (three to 12 months) and long-term (more than 12 months) duration.
[N]6.1	Intervention	What are the indications for referring people with gout to specialist services?	All outcomes are considered equally important for decision making and therefore have all been rated as critical: • health-related quality of life (e.g. as described by SF-36, Gout Assessment Questionnaire (GAQ) and the Gout Impact Scale (GIS) or other validated gout-specific HRQoL measures • pain (measured on a visual analogue scale (VAS) or numerical rating scale such as the five-point Likert scale, or reported as pain relief of 50% or greater) • joint swelling/joint inflammation • joint tenderness • frequency of flares • patient global assessment of treatment success (response to treatment) (e.g. Likert scales, visual analogue scales (VAS), numerical ratings scales (NRS)) • adverse events – (1) cardiovascular, (2) renal and (3) gastrointestinal (e.g. diarrhoea) (total adverse events will be

Evidence report	Type of review	Review questions	Outcomes

			reported if the specific types of adverse events are not reported) (cardiovascular events can include stroke and coronary artery disease)
			• adverse events and complications of gout:
			  ○ radiographic joint damage
			  ○ renal stones
			  ○ tophi
			• serum urate levels
			• admissions (hospital and A&E/urgent care)
			• GP visits
			
			Timepoints: short-term (less than three months), medium-term (three to 12 months) and long-term (more than 12 months) duration.

[O]6.2	Intervention	What is the clinical and cost effectiveness of surgical excision of tophi (deposits of monosodium urate crystals) in people with gout?	All outcomes are considered equally important for decision making and therefore have all been rated as critical: • health-related quality of life (e.g. as described by SF-36, Gout Assessment Questionnaire (GAQ) and the Gout Impact Scale (GIS) or other validated gout-specific HRQoL measures • pain (measured on a visual analogue scale (VAS) or numerical rating scale such as the five-point Likert scale, or reported as pain relief of 50% or greater) • patient global assessment of treatment success (response to treatment) (e.g. Likert scales, visual analogue scales

Evidence report	Type of review	Review questions	Outcomes
			(VAS), numerical ratings scales (NRS)) • adverse events – (1) cardiovascular, (2) renal and (3) gastrointestinal (e.g. diarrhoea) (total adverse events will be reported if the specific types of adverse events are not reported) • adverse events and complications of gout: o radiographic joint damage o tophi • Surgical complications (wound healing, infection) • serum urate levels • admissions (hospital and A&E/urgent care) • GP visits Timepoints: short-term (less than three months), medium-term (three to 12 months) and long-term (more than 12 months) duration

2.2 Searching for evidence

2.2.1 Clinical and health economics literature searches

The full 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.

Systematic literature searches were undertaken to identify all published clinical and health economic evidence relevant to the review questions. Searches were undertaken according to the parameters stipulated within the NICE guidelines manual.[10] 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, and where possible, searches were restricted to English language. All searches were updated on 06 July 2021. 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. Checking key papers were retrieved and Medline search strategies were peer reviewed by a second information specialist using a QA process based on the PRESS checklist.[5] Additional studies were added by checking reference lists of relevant systematic reviews, and those highlighted by committee members.

During the scoping stage, a search was conducted for guidelines and reports on the websites including:

- 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)
- NHS Evidence Search (www.evidence.nhs.uk).
- TRIP (www.tripdatabase.com)
- NIHR Evidence (https://evidence.nihr.ac.uk/)

Searching for unpublished literature was not undertaken

2.2.2 Supplementary evidence

The gout guideline was selected for the collaborative research project between NICE and the University of Edinburgh entitled 'Multimorbidity and clinical guidelines: using epidemiology to quantify the applicability of trial evidence to Inform guideline development'. The aim of the project is to use high-quality contemporary UK data to systematically examine differences between people with gout who are considered either eligible or ineligible for inclusion in trials. This may inform national clinical guideline development and how guidelines covering a single disease do not take into account people with comorbidities or co-prescribing, or age because they have not been included in the trials within the systematic reviews considered by committees.

A Clinical Practice Research Datalink (CPRD) dataset (UK general practice data) was extracted at a single cross-sectional date of 30/11/15 to provide the committee with data about the clinical population of people with gout registered with UK general practitioners.

The committee discussed and agreed the data on the gout population that would be informative across all questions being addressed in the guideline (Appendix 1). This comprised of a general overview of the population with gout for example age, sex, distribution, prevalence of comorbidities including CKD and osteoarthritis and percentages of gout patients with multiple comorbidities. The percentage of people being prescribed ULT medication allopurinol or febuxostat, and co-prescribing relevant to treatments being considered within the guideline (Appendix 1). In addition, example data was provided about trial eligible vs trial ineligible people with gout for trials included in evidence reviews. In particular pharmacological treatments for gout flares and urate lowering therapies for long term treatment of the condition (Appendix 1).

The data was analysed and presented to the committee by the University of Edinburgh team and this was reviewed by the committee. The committee mainly used the data to provide background information and context to the review questions addressed. Where the data was considered by the committee this is described within the committee discussion sections within individual evidence reviews. E: Evidence review which patients should be selected for ULT 4.1a and L: evidence review optimum frequency of monitoring 5.3. The data provided did not directly influence the recommendations made by the committee, but was used as supplementary information to aid discussion, particularly where the evidence was limited.

The data has also been used in two costing analyses conducted. L: Evidence review optimum frequency of monitoring 5.3 and this is described within 1.1.8, and G Evidence

review urate lowering therapies for long term management 4.2-4.4 within 1.1.9. The CPRD data was used to estimate the proportion of people treated with different types of drugs for treatment of a gout flare in order to estimate the total drug costs associated with a gout flare. Additional CPRD data for the proportion of people with gout without comorbidities was also used in the costing analysis for Evidence L to estimate the total cost of monitoring.

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 preferred study design checklist as specified in the NICE guidelines manual.[10] Prognostic studies were critically appraised using QUIPs. Qualitative studies were critically appraised using the GRADE CERQual approach for rating confidence in the body of evidence as a whole and the CASP checklist for the methodological limitations section of the quality assessment.
- 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.
  - Data from non-randomised studies were meta-analysed if appropriate and reported in GRADE profile tables.
  - Prognostic data were meta-analysed where appropriate and reported in GRADE profile tables.
  - Diagnostic data studies were meta-analysed where appropriate or presented as a range of values in adapted GRADE profile tables
  - Qualitative data were synthesised across studies and presented as summary statements with accompanying GRADE CERQual ratings for each review finding.
- 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.

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 (18 years and older) with gout in any setting.

The key population exclusion criterion was:

- People with calcium pyrophosphate crystal deposition, including pseudogout.

To ensure only drug dosages that are prescribed in the UK, at a clinically effective level are included in the drug treatment reviews, the committee pre-defined drug dosages based on the BNF, SPC or their expert opinion. The tables below show the different drugs and dosages, according to the CKD stage.

Table 2: Managing gout flares (included in review 3.1+3.2, review 4.5)

	People without CKD	People with CKD stage 3	People with CKD stage 4-5
NSAIDs			
Celecoxib (off-label)	200-400 mg daily	100-400 mg daily	100-200mg daily
Diclofenac sodium	150 mg daily	75-150 mg daily	75mg daily
Etoricoxib	120mg daily	60-120mg daily	30-60mg daily
Ibuprofen (off-label)	1.2g to 2.4g daily	0.8g to 1.6g daily	0.6g to 1.2g daily
Indomethacin	150 - 200mg daily	150 - 200mg daily	75mg to 100mg
Meloxicam (off-label)	7.5-15mg daily	7.5-15mg daily	7.5mg daily
Naproxen	750mg to 1500mg daily	500mg to 1000mg daily	250m-750mg daily
Colchicine	1-2mg per day	1-2mg per day	<=1mg, avoid if eGFR<10
Corticosteroids			
Methylprednisolone	Intra-articular: large joint (knee, ankle), 20 – 80 mg; medium joint (elbow, wrist), 10 – 40 mg; small joint (metatarsal-phalangeal joint,) 4 – 10 mg Intramuscular: 40-120mg	Intra-articular: large joint (knee, ankle), 20 – 80 mg; medium joint (elbow, wrist), 10 – 40 mg; small joint (metatarsal-phalangeal joint,) 4 – 10 mg Intramuscular: 40-120mg	Intra-articular: large joint (knee, ankle), 20 – 80 mg; medium joint (elbow, wrist), 10 – 40 mg; small joint (metatarsal-phalangeal joint,) 4 – 10 mg Intramuscular: 40-120mg

	People without CKD	People with CKD stage 3	People with CKD stage 4-5
Prednisolone	Intra-articular: large joint 10mg to 25mg Small joint: 5-10mg Intramuscular: 25-100mg Oral prednisolone: 30-40mg daily	Intra-articular: large joint 10mg to 25mg Small joint: 5-10mg Intramuscular: 25-100mg Oral prednisolone: 30-40mg daily	Intra-articular: large joint 10mg to 25mg Small joint: 5-10mg Intramuscular: 25-100mg Oral prednisolone: 30-40mg daily
Triamcinolone)	Intra-articular: large joint: 10-40mg, small joint 5-10mg Intramuscular: 40-100mg	Intra-articular: large joint: 10-40mg, small joint 5-10mg Intramuscular: 40-100mg	Intra-articular: large joint: 10-40mg, small joint 5-10mg Intramuscular: 40-100mg
IL-1 inhibitors			
Anakinra (off-label)	100mg daily	100mg alternate days to 100mg daily	100mg alternate days
Canakinumab	150mg once	150mg once	150mg once

Table 3: Long-term management of gout (urate lowering therapies) (included in review 4.2, 4.3+4.4)

	People without CKD	People with CKD stage 3	People with CKD stage 4+5
Xanthine oxidase inhibitor			
Allopurinol	100mg-900mg daily	50mg-300mg od	50-200mg od
Febuxostat	80-120mg daily	80-120mg daily	80-120mg daily
Uricosuric			
Amlodipine (off-label)	5-10mg od	5-10mg	5-10mg
Fenofibrate (off-label)	160-267mg od	67mg	Not used
Losartan (off-label)	25-100mg od	25-100mg od	25-100mg od
Vitamin C (off-label)	500mg/day	500mg/day	500mg/day
Uricase			

	People without CKD	People with CKD stage 3	People with CKD stage 4+5
Rasburicase (off-label)	0.2mg/kg	0.2mg/kg	0.2mg/kg

The committee noted the dosages for allopurinol are advised in the renal drug handbook[2] and are rarely exceeded in people with gout and CKD in clinical practice. However, the committee highlighted that randomised trials have shown that cautious gradual escalation of allopurinol to doses beyond these ranges with close monitoring can be safe and effective in people with gout and CKD. As it would be difficult to establish the escalation from the evidence, any dosages outside of those agreed were excluded.

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.

2.3.1.1 Saturation of qualitative studies

Data extraction in qualitative reviews is a thorough process. A common approach applied in systematic reviews of qualitative data is to stop extracting data once saturation has been reached. In an exploratory review, where themes are not predefined in the protocol, thematic or data extraction may be applied. For the purposes of this review, extraction of information from relevant studies was stopped when thematic and data saturation was reached, e.g. no new information was emerging for a specific theme. This includes studies that don't report any new themes additional to those already identified in the review as well as not contributing additional information to the existing themes, as well as studies which report a new theme but data from other themes in the study doesn't contribute to the existing review themes. In the latter scenario only the new theme data is extracted. These studies are not specifically excluded from the review as they nevertheless fit the criteria defined in the review protocol. Any studies for which data were not extracted due to data saturation having been reached, but that fit the inclusion criteria of the protocol, were listed in the table for studies ‘identified but not extracted due to saturation’ in an appendix to the qualitative evidence review.

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 included where there was an appropriate washout for the intervention, otherwise only the first treatment period was included. If non-randomised intervention studies were considered appropriate for inclusion (for example, where no randomised evidence was available for critical outcomes) the committee stated a priori in the protocol that either certain identified variables must be equivalent at baseline or else the analysis had to adjust for any baseline differences. The confounding variables identified for all reviews that would require adjustment in a multivariate analysis, was age and gender. Some reviews also included previous treatment (non-pharmacological and pharmacological use) as a confounder. If the study did not fulfil either criterion it was excluded. Please refer to the review protocols in each

evidence report for full details on the study design of studies selected for each review question.

For diagnostic review questions, cross-sectional studies were included. For prognostic review questions, prospective and retrospective cohort studies were included. Case–control studies were not included. Prospective cohort studies with a multivariate analysis are regarded as the gold standard for prognostic reviews 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.

Where data from non-randomised studies were included, the results for each outcome were presented separately for each study or meta-analysed if appropriate.

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)[17] software to combine the data given in all studies for each of the outcomes of interest for the review question.

Most analyses were stratified by CKD stage (stages 3, 4-5, people without CKD or stages 1-2 and mixed CKD and no-CKD), where it was thought relevant to the review question. This meant that different studies with different stage categories of CKD were not combined and analysed together. For some questions additional stratification was used. and this is documented in the individual review question protocols in each evidence report. When additional strata were used this led to substrata (for example, using 2 stratification criteria leads to 4 substrata, using 3 stratification criteria leads to 9 substrata) which were analysed separately.

For all reviews studies were also grouped according to timepoints, mostly this was short-term (less than 3 months), medium-term (3-12 months) and long-term (more than 12 months).

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:

- joint swelling/joint inflammation
- joint tenderness
- frequency of flares
- adverse events – cardiovascular, renal and gastrointestinal
- adverse events and complications of gout
- serum urate levels
- admissions (hospital and A&E)
- GP visits

The absolute risk difference was also calculated using GRADEpro[6] 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 there were zero events in both arms risk difference was calculated.

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)
- pain
- patient global assessment of treatment success (response to treatment)

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[17] 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 6.2, updated 2021) 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.[17] If the control event rate was reported this was used to generate the absolute risk difference in GRADEpro.[6] Where this was not reported and the absolute effect could not be derived, the committee would look at the relative effect.

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 the distribution of effects. Where significant heterogeneity was present, predefined subgrouping of studies was carried out for either:

- Setting (primary and secondary)
- Choice of drug (drugs within the class, based on the intervention arm only)
- Age (over 65 years; under 65 years)

If the subgroup analysis resolved heterogeneity within all of the derived subgroups, then each of the derived subgroups would have been adopted as separate outcomes (providing at least 1 study remained in each subgroup. For example, instead of the single outcome of '*pain'*, this would have been separated into 2 outcomes '*pain in people aged under 65'* and '*pain in people aged 65 and over'*. Subgrouping was not possible because there was always 1 study remaining in one subgroup, which is not enough data to prove that it is the subgroup that is resolving the heterogeneity. Assessments of potential differences in effect between subgroups would have been based on the chi-squared tests for heterogeneity statistics between subgroups. Any subgroup differences would have been interpreted with caution as separating the groups breaks the study randomisation and as such is subject to uncontrolled confounding.

As we were unable to split the studies according to predefined strategies of subgrouping we were unable to explain statistical heterogeneity within each derived subgroup, so 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 would be described narratively. This was not thought appropriate for any of the meta-analyses.

2.3.3.1.4 Complex analysis

Network meta-analysis was considered for the 4.2-4.4 review (ULT for the long-term management of gout) but was not pursued because of insufficient data available for the relevant outcomes.

2.3.3.2 Only one study in one review (Evidence review I: diet and lifestyle interventions) used a crossover design. There was no washout period so only the first period was reported. It should be noted that this may reduce the power of the study because it is halving the sample size. [1, 4, 23]Data synthesis for prognostic factor review

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, where available. Studies were only included if the confounders prespecified by the committee were adjusted for in multivariate analysis.

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. These can only be meta-analysed if they use the same confounders and are homogenous in all other aspects of the protocol. Studies with a higher risk of bias were not included.

There was no data found to be combined in meta-analyses for the prognostic review.

2.3.3.3 Data synthesis for diagnostic test accuracy reviews

2.3.3.3.1 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, and different thresholds could be used. The thresholds were prespecified by the committee including whether or not data could be pooled across a range of thresholds. Diagnostic test accuracy measures used in the analysis were: area under the receiver operating characteristics (ROC) curve (AUC), and, for different thresholds (if appropriate), sensitivity and specificity. 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 80% will only miss 20% 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 80% will only incorrectly diagnose 20% of people who do not have the condition as positive. For this guideline, sensitivity was considered important due to the consequences of a missed case of gout (false negative result), as this could lead to delay in adequate treatment, hospitalisation and surgery. However specificity was also important as people may be treated incorrectly with long-term medication if they were diagnosed with gout and did not have it (false positive).

Coupled forest plots of sensitivity and specificity with their 95% CIs across studies (at various thresholds) were produced for each test, using RevMan5.[17] 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 was conducted where appropriate, that is, when 3 or more studies were available per threshold. Test accuracy for the studies was pooled using the bivariate method for the direct estimation of summary sensitivity and specificity using a random-effects approach in WinBUGS software.[23] The advantage of this approach is that it produces summary estimates of sensitivity and specificity that account for the correlation between the 2 statistics. Other advantages of this method have been described elsewhere.[16, 21, 22] The bivariate method uses logistic regression on the true positives, true negatives, false positives and false negatives reported in the studies. Overall sensitivity and specificity and confidence regions were plotted (using methods outlined by Novielli 2010.[14]) The pooled median sensitivity and specificity and their 95% CIs were reported in the clinical evidence summary tables. For analyses with fewer than 3 studies included, the results of the study with the lower sensitivity value was reported when there were 2 studies, or reported individually for a single study.

If appropriate, to allow comparison between tests, summary ROC curves were generated for each diagnostic test from the pairs of sensitivity and specificity calculated from the 2×2 tables, selecting 1 threshold per study. A ROC plot shows true positive rate (sensitivity) as a function of false positive rate (1 minus specificity). Data were entered into RevMan5[17] and ROC curves were fitted using the Moses-Littenberg approach. In order to compare diagnostic tests, 2 or more tests were plotted on the same graph. The performance of the different diagnostic tests was then assessed by examining the summary ROC curves visually: the test that had a curve lying closest to the upper left corner (100% sensitivity and 100% specificity) was interpreted as the best test.Heterogeneity or inconsistency amongst studies was visually inspected in the forest plots and pooled diagnostic meta-analysis plots.

Area under the ROC curve (AUC) data for each study were also plotted on a graph, for each diagnostic test. The AUC describes the overall diagnostic accuracy across the full range of thresholds. The following criteria were used for evaluating AUCs:

- ≤0.50: worse than chance
- 0.50–0.60: very poor
- 0.61–0.70: poor
- 0.71–0.80: moderate
- 0.81–0.92: good
- 0.91–1.00: excellent or perfect test.

Heterogeneity or inconsistency amongst studies was visually inspected.

2.3.3.4 Data synthesis for qualitative study reviews

The main findings for each included paper were identified and thematic analysis methods were used to synthesise this information into broad overarching themes which were summarised into the main review findings. The evidence was presented in the form of a narrative summary detailing the evidence from the relevant papers and how this informed the overall review finding plus a statement on the level of confidence for that review finding. Considerable limitations and issues around relevance were listed. A summary evidence table with the succinct summary statements for each review finding was produced including the associated quality assessment.

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 'Grading of Recommendations Assessment, Development and Evaluation (GRADE) toolbox' developed by the international GRADE working group (http://www.gradeworkinggroup.org/). The software (GRADEpro[6]) 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 4.

Table 4: 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 5.

Each outcome had its risk of bias assessed within each study first using the appropriate checklist for the study design (Cochrane RoB 2 for RCTs, or ROBINS-I for non-randomised studies or ROBIS for systematic reviews). For each study, if there was no risk of bias in any domain, the risk of bias was given a rating of 0; 'no serious risk of bias'. 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. An overall rating is calculated across all studies by taking into account the weighting of studies according to study precision. For example if the most precise studies tended to each have a score of −1 for that outcome, the overall score for that outcome would tend towards −1.

Table 5: 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: • knowledge of that participant's likely prognostic characteristics, and • 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: • the experience of the placebo effect • performance in outcome measures • the level of care and attention received, and • the methods of measurement or analysis 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: • Stopping early for benefit observed in randomised trials, in particular in the absence of adequate stopping rules. • Use of unvalidated patient-reported outcome measures. • Lack of washout periods to avoid carry-over effects in crossover trials. • Recruitment bias in cluster-randomised trials.

The assessment of risk of bias differs for non-randomised intervention studies, as they are inherently at high risk of selection bias. For this reason, GRADE requires that non-randomised evidence is initially downgraded on the basis of study design, starting with a rating of −2. This accounts for selection bias and so non-randomised intervention studies are not downgraded any further on that domain. Non-randomised evidence was assessed against the remaining domains used for RCTs in Table 5, and downgraded further as appropriate.

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.

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. 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 and a random effects model used. 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 $I^2<50\%$) then each of the derived subgroups were presented separately for that forest plot (providing at least 2 studies remained in each subgroup). 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 same MIDs for imprecision were used for the overall clinical importance of outcomes, as detailed below.

Clinical importance of outcomes for decision-making

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 is the 'default' method, as follows:

- For categorical outcomes the MIDs were taken to be RRs of 0.8 and 1.25. For 'positive' outcomes such as 'patient satisfaction', the RR of 0.8 is taken as the line denoting the boundary between no clinically important effect and a clinically significant harm, whilst the RR of 1.25 is taken as the line denoting the boundary between no clinically important effect and a clinically significant benefit. For 'negative' outcomes such as 'bleeding', the opposite occurs, so the RR of 0.8 is taken as the line denoting the boundary between no clinically important effect and a clinically significant benefit, whilst the RR of 1.25 is taken as the line denoting the boundary between no clinically important effect and a clinically significant harm.
- For mortality any change was considered to be clinically important and the imprecision was assessed on the basis of the whether the confidence intervals crossed the line of no effect, that is whether the result was consistent with both benefit and 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 significant 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.Clinically significant 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, where the GC are able to specify a priority measure, the results are back-converted to a mean difference on that scale for the assessment of imprecision and clinical importance. If it is not deemed appropriate to back-convert to a single scale, 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.

For this guideline, appropriate MIDs for continuous or dichotomous outcomes which were found in the literature are detailed in the tables below:

Table 6: Continuous MIDs

Outcome measure	MID
Patient Reported Outcome (VAS) [20] Scale: 0 (worst) -100 (best)	Improvements of ≥ 10 points
Short Form-36 (SF-36) [3, 8, 18-20] Scale: 0 (worst) - 100 (best)	SF-36 suggested cut-off point: Improvements 7.5 (range 5-10) Physical component 3.75 (range 2.5-5)
Gout Impact Scale (GIS) [7-9, 19] Scale: 0 (worse) -100 (best)	Improvements of 5.2-7.6 points Gout concern overall, 7.2; unmet gout treatment need, 6.9; gout well-being during attack, 5.2; and gout concern during attack, 7.6
Short Form-6D (SF-6D) [20]	0.041

Outcome measure	MID
0 (death) - 1 (perfect health)	
Gout Assessment Questionnaire (GAQ) 1.0 [3] Scale: 0 (worst) – 100 (best)	6.5 (range: 5-8 points) (in all subscales except well-being anchored to pain frequency)
Medical Outcomes Study 20-item Short Form Health Survey (MOS 20) [3] Scale: 0 (worst) – 100 (best)	20% change in scores
Arthritis Impact Measurement Scale (AIMS) [3] Scale: 0 –10 for each section. Total health score 0 (best) – 60 (worst).	20% change in scores
Health Assessment Questionnaire Disability Index (HAQ-DI) [20] Scale: 0 (best) -3 (worse)	0.22

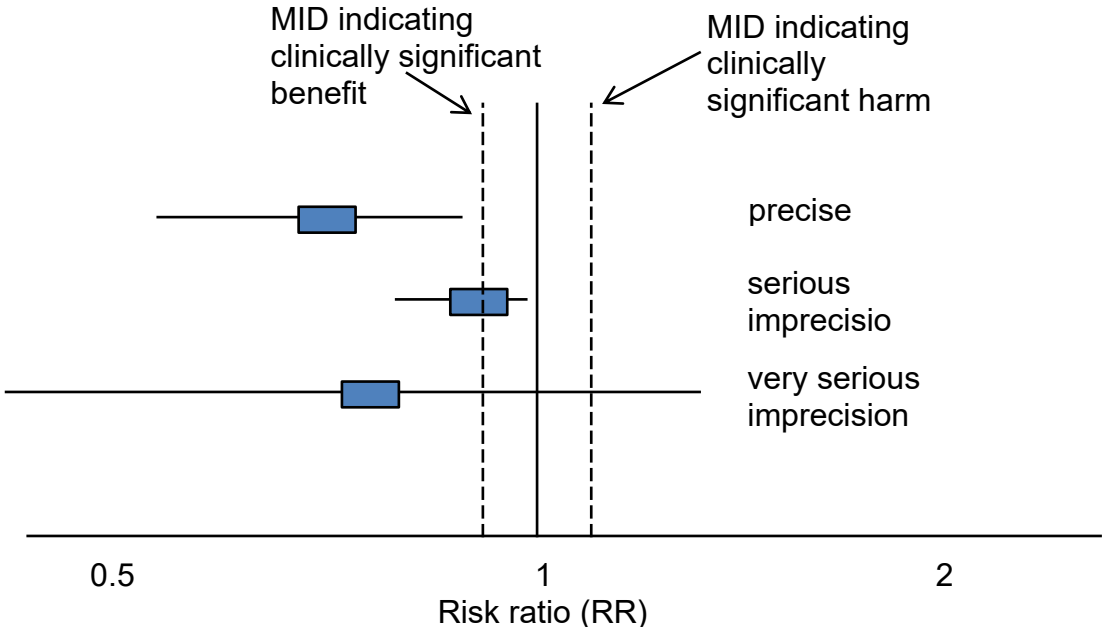
For dichotomous outcomes the GC agreed to use the National Guideline Centre standard approach:

- If at least 100 more participants per 1000 (10%) achieved the outcome of interest in the intervention group compared to the comparison group for a positive outcome, then this intervention was considered beneficial.
- The same point estimate but in the opposite direction applied for a negative outcome
- For adverse events - 50 events or more per 1000 (5%) represented clinical harm

Table 7: Dichotomous MIDs

Outcome	Number of events/1,000 patients
Frequency of flares	100 fewer per 1,000 patients = clinical benefit of intervention
Adverse events – (1) cardiovascular, (2) renal and (3) gastrointestinal (e.g. diarrhoea)	50 more per 1,000 patients = clinical harm of intervention
Renal stones	50 more per 1,000 patients = clinical harm of intervention
Tophi	50 fewer per 1,000 patients = clinical benefit of intervention
Admissions (hospital and A&E)	100 fewer per 1,000 patients = clinical benefit of intervention
GP visits	100 fewer per 1,000 patients = clinical benefit of intervention

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 8. 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 8: 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

2.3.4.2 Prognostic reviews

The risk of bias for prognostic studies was evaluated according to the QUIPS checklist, the main criteria are given in Table 5.

Table 9: Description of risk of bias criteria for prognostic studies

Risk of bias	Description of cases where the quality measure would be downgraded
Study participation	To judge selection bias (likelihood that relationship between the prognostic factor and outcome is different for participants and eligible non-participants).
Study attrition	To judge the risk of attrition bias (likelihood that relationship between prognostic factor and outcome are different for completing and non-completing participants).
Prognostic factor measurement	To judge the risk of measurement bias related to how the prognostic factor was measured (differential measurement of prognostic factor related to the baseline level of outcome).
Outcome measurement	To judge the risk of bias related to the measurement of outcome (differential measurement of outcome related to the baseline level of prognostic factor).
Study confounding	To judge the risk of bias due to confounding (i.e. the effect of the prognostic factor is distorted by another factor that is related to the prognostic factor and outcome).
Statistical analysis and reporting	To judge the risk of bias related to the statistical analysis and presentation of results.

2.3.4.2.1 Inconsistency

Inconsistency was assessed as for intervention studies.

2.3.4.2.2 Imprecision

In meta-analysed outcomes, or for non-pooled outcomes, 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 and GRADE row 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 studies, and each major limitation brought the rating down by 1 increment to a minimum grade of Very Low, as explained for interventional reviews.

2.3.4.3 Diagnostic studies

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.[10] 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 (yes/no/ unclear)	Was a consecutive or random sample of patients enrolled?	Were the index test results interpreted without knowledge of the results of the reference standard?	Is the reference standard likely to correctly classify the target condition?	Was there an appropriate interval 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 primary outcome

measures (sensitivity and specificity) using the point estimates and 95% CIs of the individual studies on the forest plots or the summary value if a diagnostic meta-analysis had been conducted. Particular attention was placed on values above or below 50% (diagnosis based on chance alone) and the threshold set by the committee (the threshold above which it would be acceptable to recommend a test). The committee set a threshold of 80% as an acceptable level to recommend at test. The evidence was downgraded by 1 increment if the individual studies varied across 2 areas (50–80% and 80–100%) and by 2 increments if the individual studies varied across 3 areas (0–50%, 50–80% and 80–100%).

2.3.4.3.2 Imprecision

The judgement of precision was based on visual inspection of the confidence region around the summary sensitivity and specificity point from the diagnostic meta-analysis, if a diagnostic meta-analysis was conducted. Where a diagnostic meta-analysis was not conducted, imprecision was assessed according to the range of point estimates or, if only one study contributed to the evidence, the 95% CI around the single study. The evidence was downgraded by one increment if the 95% confidence interval crossed one clinical decision threshold and by two increments if it crossed two clinical decision thresholds. The GC set the thresholds for sensitivity and specificity as 50% (no better than chance) and 80% (threshold to recommend a test). Where there were 3 studies, imprecision was assessed on confidence intervals produced by WinBUGS; where there were 2 studies, the results of the study with the lower sensitivity value was reported when there were 2 studies, or reported individually for a single study.

2.3.4.3.3 Overall grading

Quality rating started at High for prospective and retrospective cross-sectional studies, 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. This was presented in a GRADE profile.

2.3.4.4 Qualitative reviews

Review findings from the included qualitative studies were evaluated and presented using the 'Confidence in the Evidence from Reviews of Qualitative Research' (CERQual) Approach developed by the GRADE-CERQual Project Group, a subgroup of the GRADE Working Group.

The CERQual Approach assesses the extent to which a review finding is a reasonable representation of the phenomenon of interest (the focus of the review question). Each review finding was assessed for each of the 4 quality elements listed and defined below in Table 10.

Table 10: Description of quality elements in GRADE-CERQual for qualitative studies

Quality element	Description
Methodological limitations	The extent of problems in the design or conduct of the included studies that could decrease the confidence that the review finding is a reasonable representation of the phenomenon of interest. Assessed at the study level using the CASP checklist.
Coherence	The extent to which the reviewer is able to identify a clear pattern across the studies included in the review.
Relevance	The extent to which the body of evidence from the included studies is applicable to the context (study population, phenomenon of interest, setting) specified in the protocol.
Adequacy	The degree of the confidence that the review finding is being supported by sufficient data. This is an overall determination of the richness (depth of analysis) and quantity of the evidence supporting a review finding or theme.

Details of how the 4 quality elements (methodological limitations, coherence, relevance and adequacy) were appraised for each review finding are given below.

2.3.4.4.1 Methodological limitations

Each review finding had its methodological limitations assessed within each study first using the CASP checklist. Based on the degree of methodological limitations studies were evaluated as having minor, moderate or severe limitations. A summary of the domains and questions covered is given below.

Table 11: Description of limitations assessed in the CASP checklist for qualitative studies

Domain	Aspects considered
Are the results valid?	• Was there a clear statement of the aims of the research? • Is qualitative methodology appropriate? • Was the research design appropriate to address the aims of the research? • Was the recruitment strategy appropriate to the aims of the research? • Was the data collected in a way that addressed the research issue? • Has the relationship between researcher and participants been adequately considered?
What are the results?	• Have ethical issues been taken into consideration? • Was the data analysis sufficiently rigorous? • Is there a clear statement of findings?
Will the results help locally?	How valuable is the research?

The overall assessment of the methodological limitations of the evidence was based on the primary studies contributing to the review finding. The relative contribution of each study to the overall review finding and of the type of methodological limitation(s) were taken into account when giving an overall rating.

2.3.4.4.2 Coherence

Coherence is the extent to which the reviewer is able to identify a clear pattern across the studies included in the review, and if there is variation present (contrasting or disconfirming data) whether this variation is explained by the contributing study authors. For example, if a review finding in 1 study does not support the main finding and there is no plausible explanation for this variation, or if there is ambiguity in the descriptions in the primary data, then the confidence that the main finding reasonably reflects the phenomenon of interest is decreased.

2.3.4.4.3 Relevance

Relevance is the extent to which the body of evidence from the included studies is applicable to the context (study population, phenomenon of interest, setting) specified in the protocol. As such, relevance is dependent on the individual review and discussed with the guideline committee.

2.3.4.4.4 Adequacy

The judgement of adequacy is based on the confidence of the finding being supported by sufficient data. This is an overall determination of the richness (depth of analysis) and

quantity of the evidence supporting a review finding or theme. Rich data provide sufficient detail to gain an understanding of the theme or review finding, whereas thin data do not provide enough detail for an adequate understanding. Quantity of data is the second pillar of the assessment of adequacy. For review findings that are only supported by 1 study or data from only a small number of participants, the confidence that the review finding reasonable represents the phenomenon of interest might be decreased. As with richness of data, quantity of data is review dependent. Based on the overall judgement of adequacy, a rating of no concerns, minor concerns, or substantial concerns about adequacy was given.

2.3.4.4.5 Overall judgement of the level of confidence for a review finding

GRADE-CERQual is used to assess the body of evidence as a whole through a confidence rating representing the extent to which a review finding is a reasonable representation of the phenomenon of interest. The 4 components (methodological limitations, coherence, relevance and adequacy) are used in combination to form an overall judgement. GRADE-CERQual uses 4 levels of confidence: high, moderate, low and very low confidence. The significance of these overall ratings is explained in Table 12. Each review finding starts at a high level of confidence and is downgraded based on the concerns identified in any 1 or more of the 4 components. Quality assessment of qualitative reviews is a subjective judgement by the reviewer based on the concerns that have been noted. A detailed explanation of how such a judgement had been made was included in the narrative summary.

Table 12: Overall level of confidence for a review finding in GRADE-CERQual

Level	Description
High confidence	It is highly likely that the review finding is a reasonable representation of the phenomenon of interest.
Moderate confidence	It is likely that the review finding is a reasonable representation of the phenomenon of interest.
Low confidence	It is possible that the review finding is a reasonable representation of the phenomenon of interest.
Very low confidence	It is not clear whether the review finding is a reasonable representation of the phenomenon of interest.

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[6] 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. The committee considered for most of the outcomes in the intervention reviews that if at least 100 more participants per 1000 (10%) achieved the outcome of interest in the intervention group compared to the comparison group for a positive outcome then this intervention was considered beneficial. The same point estimate but in the opposite direction applied for a negative outcome. For the critical outcome of mortality any reduction represented a clinical benefit. For adverse events 50 events or more per 1000 (5%) represented clinical harm. For continuous outcomes if the mean difference was greater than the minimally important difference (MID) then this represented a clinical benefit or harm. For outcomes such as mortality any reduction or increase was considered to be clinically important. Individual MIDs can be found in Table 6 and Table 7.

This assessment was carried out by the committee for each critical outcome, and an evidence summary table was produced to compile the committee's assessments of clinical importance per outcome, alongside the evidence quality and the uncertainty in the effect estimate (imprecision).

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.[10]

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.
- Undertook costing analyses in priority areas.

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.[10]
- 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.

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 2005 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. Where exclusions occurred on this basis, this is noted in the relevant evidence report.

For more details about the assessment of applicability and methodological quality see Table 13 below and the economic evaluation checklist (appendix H of the NICE guidelines manual[10]) 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.[10] 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 13 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.[15]

Table 13: 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)] • 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. • Partially applicable – the study fails to meet 1 or more applicability criteria, and this could change the conclusions about cost effectiveness. • Not applicable – the study fails to meet 1 or more of the applicability criteria, and this is likely to change the conclusions about cost effectiveness. Such studies would usually be excluded from the review.
Limitations	An assessment of methodological quality of the study:[(a)] • 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. • Potentially serious limitations – the study fails to meet 1 or more quality criteria, and this could change the conclusions about cost effectiveness. • 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.

Item	Description
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[10]

2.4.2 Undertaking new health economic analysis

As well as reviewing the published health economic literature for each review question, as described above, new health economic analysis was undertaken by the health economist in selected areas. Priority areas for new analysis were agreed by the committee after formation of the review questions and consideration of the existing health economic evidence.

The committee identified which ULTs should be prescribed as a first- and second-line therapy, and the optimum frequency of monitoring once people achieve target serum urate levels as the highest priority areas for original health economic modelling.

After review of the clinical evidence for which ULTs should be offered as first and second line therapy, and discussion with the committee, it was concluded an original health economic model would unlikely reduce the uncertainty of the cost effectiveness of allopurinol and febuxostat due to a lack of additional clinical evidence published since the previous Technology Appraisal assessing the cost effectiveness of febuxostat. Although additional clinical evidence was available reporting the proportion of people receiving higher doses of 300mg allopurinol the results for people achieving target serum urate levels were not stratified by dose. Therefore, any further modelling would likely be a duplication of the existing economic models and their associated limitations such as the lack of evidence for the use of allopurinol at doses greater than 300mg, or that similar model assumptions would need to be made in terms of linking sUA to probability of gout flares (based on unpublished data in Beard 2013). In addition, given that the cost of febuxostat 80mg and allopurinol at doses greater than 300mg are so similar, it is likely that the results of any further modelling would be sensitive to any model assumptions made with regard to the effectiveness of allopurinol at doses greater than 300mg. Given these concerns, it was agreed to undertake a costing analysis rather than a cost-utility analysis to aid the committee in their consideration of the cost effectiveness of allopurinol and febuxostat. This analysis determined which ULT (allopurinol and febuxostat) was the least and most costly intervention over a one-year time horizon with a number of different scenarios to account for uncertainty.

No clinical or health economic evidence was identified for the optimum frequency of monitoring. The committee acknowledged that monitoring people once they have achieved target serum urate levels could improve people's medication adherence and health outcomes because currently people may not realise, they have deviated from target serum urate levels until they experience more frequent or severe flares. In which case, it may be more difficult and resource intensive to reobtain target serum urate levels compared to if identified at a monitoring appointment. Therefore, a costing analysis was undertaken to assess the number of gout flares needed be avoided per person per year for annual monitoring to break even.

The following general principles were adhered to in developing the costing analyses:

- Methods were consistent with the NICE reference case for interventions with health outcomes in NHS settings.[10, 13]
- The committee was involved in the design of the analyses, selection of inputs and interpretation of the results.
- For the which ULT costing analysis, inputs were based on the systematic review of the clinical literature supplemented with other published data sources where possible.

- When published data was not available committee expert opinion was used. Data from the Multimorbidity and clinical guidelines project was also used in the costing analyses.
- Inputs and assumptions were reported fully and transparently.
- The results were subject to sensitivity analysis and limitations were discussed.
- The analysis was peer-reviewed by another health economist at the NGC.

Full methods and results of the costing analyses can be found in the corresponding evidence reviews (Evidence review: G, Urate lowering therapies for the long-term management of gout and Evidence review L: Optimum frequency of monitoring).

2.4.3 Cost-effectiveness criteria

NICE sets out the principles that committees should consider when judging whether an intervention offers good value for money.[10-12] 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 factors set out in NICE methods manuals.[10]

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.

2.4.4 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.

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–O).
- 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 and summary ROC curves (in appendices to the relevant evidence reports).
- A description of the methods and results of the costing analyses undertaken for the guideline (included in their respective 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. The committee also considered whether the uncertainty was sufficient to justify delaying making a recommendation to await further research, taking into account the potential harm of failing to make a clear recommendation (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 patients 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 patients.

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.
- The strength of the recommendation (for example the word 'offer' was used for strong recommendations and 'consider' for weaker recommendations).
- The involvement of patients (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[10]).

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[10], 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.

3 Additional information

4 Acronyms and abbreviations

Acronym or abbreviation	Description
AUC	Area under the curve
BNF	British National Formulary
CI	Confidence interval
CKD	Chronic Kidney Disease
COMET	Core Outcome Measures in Effectiveness Trials
CUA	Cost-utility analysis
EQ-5D	European Quality of Life Five Dimension
GC	Guideline committee
GRADE	Grading of Recommendations Assessment, Development and Evaluation
GRADE-CERQual	Grading of Recommendations Assessment, Development and Evaluation in qualitative research
HAQ-DI	Health Assessment Quality of life Disability Index
HCP	Health care practitioner
HR	Hazard ratio
HRQOL	Health-related quality of life
ICER	Incremental cost-effectiveness ratio
IMMPACT	Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials
NGC	National Guideline Centre
NICE	National Institute for Health and Care Excellence
NSAID	Non-steroidal anti-inflammatory drug
OECD	Organisation for Economic Co-operation and Development
OR	Odds ratio
QALY	Quality-adjusted life year
QoL	Quality of life
RCT	Randomised controlled trial
ROC	Receiver operating characteristic
RR	Risk ratio
SF-36	36-Item Short Form Health Survey
SUA	Serum urate acid
ULT	Urate-lowering therapy
VAS	Visual analogue scale

5 Glossary

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

5.1 Guideline-specific terms

Term	Definition
Chronic gouty arthritis	Persistent joint inflammation caused by monosodium urate crystals.
Chronic kidney disease	Abnormalities of kidney function or structure present for more than 3 months, with implications for health. This includes all people with markers of kidney damage and those with a glomerular filtration rate (GFR) of less than 60 ml/min/1.73m^2 on at least 2 occasions separated by a period of at least 90 days (with or without markers of kidney damage).
Gout	The most common type of inflammatory arthritis.
Gout flare	A clinically evident episode of acute inflammation affecting a joint caused by monosodium urate crystals.
Hyperuricaemia	Elevated blood urate concentration.
Monosodium urate crystals	Pathogenic crystals which form as a result of an elevated blood urate concentration and deposit in and round joints causing gout.
Prophylaxis	A drug (non-steroidal anti-inflammatory drug, colchicine, corticosteroid or IL-1 inhibitor) given to prevent gout flares when starting urate-lowering therapy.
Serum urate	A naturally occurring substance in the body. Elevated levels can lead to the formation of monosodium urate crystals and gout.
Tophus/tophi	A discrete lump of monosodium urate crystals, usually overlying a joint (singular tophus, plural tophi).
Treat-to-target	A treatment approach which involves gradually increasing the dose of urate-lowering therapy with the aim of lowering the serum urate level below a pre-specified target level.
Urate-lowering therapy	Drug therapy (most commonly allopurinol or febuxostat) which lowers the serum urate level.

5.2 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 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.

Term	Definition
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.
Bayesian analysis	A method of statistics, where a statistic is estimated by combining established information or belief (the 'prior') with new evidence (the 'likelihood') to give a revised estimate (the 'posterior').
Before-and-after study	A study that investigates the effects of an intervention by measuring particular characteristics of a population both before and after taking the intervention, and assessing any change that occurs.
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. 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. 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 comparison (control) group of patients.
Clinical efficacy	The extent to which an intervention is active when studied under controlled research conditions.

Term	Definition
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. 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. 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. 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. 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 to a question. Formal consensus methods include Delphi and nominal group techniques.

Term	Definition
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. 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.
Credible interval (CrI)	The Bayesian equivalent of a confidence interval.
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

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. 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 (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. For example, if the absolute risk reduction is shown to be 5% and it is the outcome of interest, the effect size is 5%. 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.
Extended dominance	If Option A is both more clinically effective than Option B and has a lower cost per unit of effect, when both are compared with a do-nothing alternative then Option A is said to have extended dominance over Option B. Option A is therefore cost effective and should be preferred, other things remaining equal.
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.

Term	Definition
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 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.
Hazard ratio	The hazard or chance of an event occurring in the treatment arm of a study as a ratio of the chance of an event occurring in the control arm over time.
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 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 analysis	The analysis of additional costs and additional clinical outcomes with different interventions.
Incremental cost	The extra cost linked to using one test or treatment rather than another. Or the additional cost of doing a test or providing a treatment more frequently.
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.
Incremental net benefit (INB)	The value (usually in monetary terms) of an intervention net of its cost compared with a comparator intervention. The INB can be calculated for a given cost-effectiveness (willingness to pay) threshold. If the threshold is £20,000 per QALY gained then the INB is calculated as: (£20,000 × QALYs gained) − Incremental cost.
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.

Term	Definition
Intraoperative	The period of time during a surgical procedure.
Kappa statistic	A statistical measure of inter-rater agreement that takes into account the agreement occurring by chance.
Length of stay	The total number of days a participant stays in hospital.
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).
Long-term care	Residential care in a home that may include skilled nursing care and help with everyday activities. This includes nursing homes and residential homes.
Logistic regression or Logit model	In statistics, logistic regression is a type of analysis used for predicting the outcome of a binary dependent variable based on one or more predictor variables. It can be used to estimate the log of the odds (known as the 'logit').
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 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)
Net monetary benefit (NMB)	The value in monetary terms of an intervention net of its cost. The NMB can be calculated for a given cost-effectiveness threshold. If the threshold is £20,000 per QALY gained then the NMB for an intervention is calculated as: (£20,000 × mean QALYs) − mean cost. The most preferable option (that is, the most clinically effective option to have an ICER below the threshold selected) will be the treatment with the highest NMB.
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. 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.

Term	Definition
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. 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. 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. 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. 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 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. 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.

Term	Definition
	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.
Perioperative	The period from admission through surgery until discharge, encompassing the preoperative and postoperative periods.
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.
Posterior distribution	In Bayesian statistics this is the probability distribution for a statistic based after combining established information or belief (the prior) with new evidence (the likelihood).
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)
Postoperative	Pertaining to the period after patients leave the operating theatre, following surgery.
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.
Preoperative	The period before surgery commences.
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.
Prior distribution	In Bayesian statistics this is the probability distribution for a statistic based on previous evidence or belief.
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 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.

Term	Definition
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. 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.
RCT	See 'Randomised controlled trial'.
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). 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

Term	Definition
	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: a) The characteristics of the people selected for a study differ from the wider population from which they have been drawn, or 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. 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'). 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. 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'). 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. 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. 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. Threshold sensitivity analysis: the critical value of parameters above or below which the conclusions of the study will change are identified. 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 non-cases correctly diagnosed as non-cases. See related term 'Sensitivity'.

Term	Definition
	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: • manufacturers of drugs or equipment • national patient and carer organisations • NHS organisations • organisations representing healthcare professionals.
State transition model	See Markov model
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.
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).

References

1. Ades AE, Caldwell DM, Reken S, Welton NJ, Sutton AJ, Dias S. Evidence synthesis for decision making 7: a reviewer's checklist. Medical Decision Making. 2013; 33(5):679-691

2. Ashley C, Dunleavy A, Cunnningham J, (Eds). The Renal Drug Handbook: The Ultimate Prescribing Guide for Renal Practitioners (5th ed.). CRC Press. 2018.

3. Chandratre P, Roddy E, Clarson L, Richardson J, Hider SL, Mallen CD. Health-related quality of life in gout: a systematic review. Rheumatology. 2013; 52(11):2031-2040

4. Dias S, Sutton AJ, Welton NJ, Ades AE. Evidence synthesis for decision making 3: heterogeneity--subgroups, meta-regression, bias, and bias-adjustment. Medical Decision Making. 2013; 33(5):618-640

5. Goldfien R, Pressman A, Jacobson A, Ng M, Avins A. A pharmacist-staffed, virtual gout management clinic for achieving target serum uric acid levels: A randomized clinical trial. Permanente Journal. 2016; 20(3):15-234

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

7. Khanna D, Sarkin AJ, Khanna PP, Shieh MM, Kavanaugh AF, Terkeltaub RA et al. Minimally important differences of the gout impact scale in a randomized controlled trial. Rheumatology. 2011; 50(7):1331-1336

8. Khanna PP, Perez-Ruiz F, Maranian P, Khanna D. Long-term therapy for chronic gout results in clinically important improvements in the health-related quality of life: short form-36 is responsive to change in chronic gout. Rheumatology. 2011; 50(4):740-745

9. Khanna PP, Shiozawa A, Walker V, Bancroft T, Essoi B, Akhras KS et al. Health-related quality of life and treatment satisfaction in patients with gout: results from a cross-sectional study in a managed care setting. Patient preference & adherence. 2015; 9:971-981

10. National Institute for Health and Care Excellence. Developing NICE guidelines: the manual [updated October 2020]. London. National Institute for Health and Care Excellence, 2014. Available from: http://www.nice.org.uk/article/PMG20/chapter/1%20Introduction%20and%20overview

11. National Institute for Health and Care Excellence. The NICE Charter. 2020. Available from: https://www.nice.org.uk/about/who-we-are/our-charter Last accessed: 10/03/2020.

12. National Institute for Health and Care Excellence. The principles that guide the development of NICE guidance and standards. 2020. Available from: https://www.nice.org.uk/about/who-we-are/our-principles Last accessed: 10/03/2020.

13. National Institute for Health and Clinical Excellence. Guide to the methods of technology appraisal 2013. London. National Institute for Health and Clinical Excellence, 2013. Available from: http://publications.nice.org.uk/pmg9

14. Novielli N, Cooper NJ, Abrams KR, Sutton AJ. How is evidence on test performance synthesized for economic decision models of diagnostic tests? A systematic appraisal

of Health Technology Assessments in the UK since 1997. Value in Health. 2010; 13(8):952-957

15. Organisation for Economic Co-operation and Development (OECD). Purchasing power parities (PPP). 2012. Available from: http://www.oecd.org/std/ppp Last accessed: September 2021.

16. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. Journal of Clinical Epidemiology. 2005; 58(10):982-990

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

18. Singh JA, Bharat A, Khanna D, Aquino-Beaton C, Persselin JE, Duffy E et al. Racial differences in health-related quality of life and functional ability in patients with gout. Rheumatology. 2017; 56(1):103-112

19. Singh JA, Strand V. Gout is associated with more comorbidities, poorer health-related quality of life and higher healthcare utilisation in US veterans. Annals of the Rheumatic Diseases. 2008; 67(9):1310-1316

20. Strand V, Khanna D, Singh JA, Forsythe A, Edwards NL. Improved health-related quality of life and physical function in patients with refractory chronic gout following treatment with pegloticase: evidence from phase III randomized controlled trials. Journal of Rheumatology. 2012; 39(7):1450-1457

21. Van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. Statistics in Medicine. 2002; 21(4):589-624

22. Van Houwelingen HC, Zwinderman KH, Stijnen T. A bivariate approach to meta-analysis. Statistics in Medicine. 1993; 12(24):2273-2284

23. WinBUGS [Computer programme] version 1.4. Cambridge. MRC Biostatistics Unit University of Cambridge, 2015. Available from: http://www.mrc-bsu.cam.ac.uk/software/bugs/the-bugs-project-winbugs/