National Institute for Health and Care Excellence

Draft for Consultation

Brain tumours (primary) and brain metastases in adults

Methods

NICE guideline <number>

Supplementary Material C

January 2018

Draft for Consultation

These evidence reviews were developed by the National Guideline Alliance, hosted by the Royal College of Obstetricians and Gynaecologists

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/or their carer or guardian.

Local commissioners and/or 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 All NICE guidance is subject to regular review and may be updated or withdrawn.

Copyright

ISBN:

Contents

Development of the guideline

Remit

The National Institute for Health and Care Excellence (NICE) received the remit for this guideline from the Department of Health. NICE commissioned the National Guideline Alliance (NGA) to produce the guideline.

The remit for this guideline is to develop a clinical guideline on the management of primary brain tumours and brain metastases in adults.

The scope for this guideline is provided on the NICE website.

What this guideline covers

Groups that will be covered

1. Adults (18 and over) with radiologically identified glioma, meningioma, or 1 or more brain metastases.
2. Adults with any type of primary brain tumour or brain metastases who might need assessment for neurological rehabilitation.

Settings that will be covered

1. All settings in which NHS care is provided.
2. Shared care, including social services.

Key areas that will be covered

1. Diagnosing radiologically identified glioma, meningioma and brain metastases.
2. Managing glioma.
3. Managing meningioma.
4. Managing brain metastases.
5. Follow-up care after treatment for glioma, meningioma or brain metastases.
6. Referring adults with primary brain tumours or brain metastases for neurological rehabilitation assessment.

What this guideline does not cover

1. Identifying people in primary care with suspected primary brain tumours or cerebral metastases and referring them to secondary care. This is already covered in NICE's guideline on suspected cancer: recognition and referral.
2. The following (non-exhaustive) list of tumour types:
    - neuronal and mixed neuronal-glial tumours
    - tumours of the pineal region
    - embryonal tumours
    - tumours of the cranial and paraspinal nerves
    - melanocytic tumours

- lymphomas
- mesenchymal, histiocytic, germ cell, sellar originating and choroid plexus tumours.

Methods

This guideline was developed using the methods described in the 2014 NICE guidelines manual.

Declarations of interest were recorded according to the 2014 NICE conflicts of interest policy.

Developing the review questions and outcomes

The 18 review questions developed for this guideline were based on the key areas identified in the guideline scope. They were drafted by the NGA, and refined and validated by the guideline committee. These questions are outlined in Table 1.

The review questions were based on the following frameworks:

- population, intervention, comparator and outcome (PICO) for reviews of interventions
- reviews of diagnostic test accuracy – using population, diagnostic test (index tests), reference standard and target condition
- qualitative reviews – using population, area of interest and themes of interest.

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

Description of review questions

Table 1: Description of review questions

Chapter or section from the scope	Location in Evidence Reports	Type of review	Review question	Outcomes
Diagnosing radiologically identified glioma, meningioma and brain metastases	Evidence Report A (glioma) and Evidence Report B (meningioma)	Diagnostic	1a - What is the most effective imaging strategy in newly diagnosed glioma and meningioma?	Critical: • health-related quality of life (especially anxiety) • diagnostic accuracy, including: ○ sensitivity ○ specificity ○ likelihood ratios
Diagnosing radiologically identified glioma, meningioma and brain metastases	Evidence Report C	Diagnostic[1]	1b - What is the most appropriate diagnostic imaging for patients being considered for focal treatment of their brain metastases?	Critical: • number of metastases If the critical outcome is reported, the following outcomes will be also considered:

Chapter or section from the scope	Location in Evidence Reports	Type of review	Review question	Outcomes

				Critical:
				• overall survival.
				• progression-free survival:
				  ○ local control (site of metastasis)
				  ○ intracranial control (recurrence elsewhere in the brain)
				• health-related quality of life
				
				Important:
				• cognitive function:
				  ○ neurological function
				  ○ Karnofsky performance status (or WHO or ECOG)
				• Neurological Function Scale
				• treatment-related morbidity:
				  ○ radionecrosis
				  ○ oedema
				  ○ postoperative infection
				  ○ stroke

Diagnosing radiologically identified glioma, meningioma and brain metastases	Evidence Report A	Intervention	1c - What is the optimal timing and extent of initial surgery for suspected low-grade glioma?	Critical: • progression-free survival • epilepsy/seizure control • neurological function ○ Neurological Function Scale or NIH stroke scale Important: • overall survival • time to tumour transformation (from low-grade to high-grade) • health-related quality of life.

Chapter or section from the scope	Location in Evidence Reports	Type of review	Review question	Outcomes
Diagnosing radiologically identified glioma, meningioma and brain metastases	Evidence Report A	Prognostic	1d - What are the most useful molecular markers to determine prognosis /guide treatment for gliomas?	Critical: • overall survival • progression-free survival For BRAF v600e mutation group only: • response to BRAF inhibitors (vemurafenib, daburafenib, tremetanib)
Managing glioma	Evidence Report A	Intervention	2a - What is the optimal management (observation, surgery, radiotherapy, chemotherapy or combinations of these) for histologically proven low grade glioma?	Critical: • overall survival • cognitive function • neurological function (as measured by the Neurological Function Scale or NIH stroke scale) Important: • health-related quality of life • progression free survival • epilepsy/seizure control • grade 3 or 4 late toxicity (after 3 months)

Managing glioma	Evidence Report A	Intervention	2b - What is the most effective method for optimising maximal safe resection of glioma (for example with 5-ALA, awake craniotomy, intraoperative ultrasound, intraoperative MRI)?	Critical:
				• overall survival.
				• gross total resection margins (as determined by post-operative MRI)
				• progression-free survival
				• neurological function
				  ○ Karnofsky performance status (KPS)
				  ○ Neurological Function Scale
				  ○ language
				
				Important:
				• treatment-related mortality
				• treatment-related morbidity:
				  ○ wound infection
				• length of surgery

Chapter or section from the scope	Location in Evidence Reports	Type of review	Review question	Outcomes
Managing glioma	Evidence Report A	Intervention	2c - Following surgery, what is the optimal management (radiotherapy, chemotherapy, combinations of these, or other therapies such as metformin or tumour-treating fields) of initial high-grade glioma?	Critical: • overall survival • cognitive function • neurological function (as measured by the Neurological Function Scale or NIH stroke scale) Important: • health-related quality of life • progression free survival • epilepsy/seizure control • grade 3 or 4 late toxicity (after 3 months)
Managing glioma	Evidence Report A	Intervention	2d - What is the optimal management (surgery, radiotherapy, chemotherapy, combinations of these, or other therapies such as metformin or tumour-treating fields) of recurrent high-grade glioma?	Critical: • overall survival. • progression-free survival / time to progression • health related quality of life Important: • neurological adverse events • wound infections • RTOG grade 3 and/or 4 toxicity • CTCAE grade 3 and/or 4 toxicity • fatigue (somnolence) • cognitive function

Managing meningioma	Evidence Report B	Intervention	3a - Which adults with inoperable or incompletely excised or recurrent meningioma should be offered radiotherapy?	Critical:
				• overall survival
				• progression-free survival
				• cognitive function
				• neurological function:
				  ○ cranial neuropathy (e.g. optic neuropathy)
				Important:
				• treatment-related morbidity:
				  ○ radionecrosis
				  ○ oedema
				  ○ stroke

Chapter or section from the scope	Location in Evidence Reports	Type of review	Review question	Outcomes
				○ second malignancy ○ pituitary dysfunction ○ epilepsy/seizures • health-related quality of life
Managing meningioma	Evidence Report B	Intervention	3b - Which technique should be used for adults with meningioma who require radiotherapy?	• <u>Critical:</u> ○ progression-free survival/ local control ○ Karnofsky performance status ○ steroid (for example dexamethasone) use (duration and dose) • <u>Important:</u> ○ health-related quality of life ○ Neurological Function Scale ○ cognitive function
Managing brain metastases	Evidence Report C	Intervention	4a - What is the most effective intracranial treatment (surgery, stereotactic radiotherapy, whole-brain radiotherapy or combinations of these) for a single brain metastasis?	<u>Critical:</u> • survival • progression-free survival • local control • recurrence • quality of life <u>Important:</u> • cognitive function • neurological function • treatment related morbidity
Managing brain metastases	Evidence Report C	Intervention	4b - What is the most effective intracranial treatment (surgery, stereotactic radiotherapy, whole-brain radiotherapy, combinations of these, or best supportive care) for multiple brain metastases?	<u>Critical:</u> • overall survival. • progression-free survival ○ local control (site of metastasis) ○ intracranial control (recurrence elsewhere in the brain) • health-related quality of life <u>Important:</u>

Chapter or section from the scope	Location in Evidence Reports	Type of review	Review question	Outcomes

				• cognitive function.
				• neurological function
				  ○ Karnofsky performance status (or WHO or ECOG)
				  ○ Neurological Function Scale
				• treatment-related morbidity.
				  ○ radionecrosis
				  ○ oedema
				  ○ postoperative infection
				  ○ stroke

Managing brain metastases	Evidence Report C	Intervention	4c - What is the most effective intracranial treatment (surgery, stereotactic radiotherapy, whole brain radiotherapy or combinations of these) for a mixed population of single and multiple brain metastases?	<u>Critical:</u>
				• overall survival.
				• progression-free survival
				  ○ local control (site of metastasis)
				  ○ intracranial control (recurrence elsewhere in the brain)
				• health-related quality of life
				
				<u>Important:</u>
				• cognitive function.
				• neurological function
				  ○ Karnofsky performance status (or WHO or ECOG)
				  ○ Neurological Function Scale
				• treatment-related morbidity.
				  ○ radionecrosis
				  ○ oedema
				  ○ postoperative infection
				  ○ stroke

Follow-up care after treatment for glioma, meningioma	Evidence Report A	Intervention	5a - What is the most effective follow-up protocol (including duration, frequency and tests) to detect	<u>Critical:</u> • treatment for recurrence • overall survival. • cognition

Chapter or section from the scope	Location in Evidence Reports	Type of review	Review question	Outcomes
or brain metastases			recurrence after treatment for glioma?	• symptomatic versus asymptomatic presentation Important: • health-related quality of life ○ neurological outcomes ○ seizures
Follow-up care after treatment for glioma, meningioma or brain metastases.	Evidence Report B	Intervention	5b - What is the most effective follow-up protocol (including duration, frequency and tests) to detect recurrence after treatment for meningioma?	Critical: • treatment for recurrence • overall survival. • cognition • symptomatic versus asymptomatic presentation Important: • health-related quality of life ○ neurological outcome ○ seizures
Follow-up care after treatment for glioma, meningioma or brain metastases.	Evidence Report D	Intervention	5c - What is the most effective follow-up protocol (including duration, frequency and tests) to detect intracranial recurrence after treatment for brain metastases?	Critical: • treatment for recurrence • overall survival. • cognition • symptomatic versus asymptomatic presentation Important: • health-related quality of life ○ neurological outcomes ○ seizures
Follow-up care after treatment for glioma, meningioma or brain metastases.	Evidence Report D	Intervention	5d - What is the most effective surveillance protocol (including no surveillance) for detecting late effects of treatment for glioma, meningioma or brain metastases?	Critical: • stage and incidence of late effects (occurring from 12 months after treatment onwards): ○ stroke ○ secondary cancer/tumour (in brain and body) ○ visual loss and cataract ○ hypopituitarism ○ neurocognitive decline

Chapter or section from the scope	Location in Evidence Reports	Type of review	Review question	Outcomes
				○ radio necrosis • severity of late effects: ○ stroke ○ secondary cancer/tumour (in brain and body) ○ visual loss and cataract ○ hypopituitarism ○ neurocognitive decline ○ radio necrosis • treatment of late effects: ○ stroke ○ secondary cancer/tumour (in brain and body) ○ visual loss and cataract ○ hypopituitarism ○ neurocognitive decline ○ radio necrosis
Follow-up care after treatment for glioma, meningioma or brain metastases	Evidence Report D	Qualitative	5e - What are the health and social care support needs of people with brain tumours (primary) and brain metastases and their families and carers?	Themes occurring in the context of health or social care support required by a person with a brain tumour and the family or carer of a person with a brain tumour
Referring adults with primary brain tumours or brain metastases for neurological rehabilitation assessment	Evidence Report D	Qualitative	6a - What are the facilitators and barriers to providing appropriate neurological rehabilitation assessment in people with brain tumours (primary) and brain metastases?	Themes occurring in the context of health or social care support required by a person with a brain tumour and the family or carer of a person with a brain tumour

1 *5-ALA 5-amino-levulinic acid; BRAF proto-oncogene b-raf / v-raf murine sarcoma viral oncogene*
2 *homolog b; CTCAE common terminology criteria for adverse events; ECOG Eastern Cooperative*
3 *Oncology Group; KPS Karnofsky performance status; MRI magnetic resonance imaging; NIH National*

Institute for Health; RTOG radiation therapy oncology group; WHO World Health Organization

[1] While this is classified as a diagnostic review, the outcomes to be evaluated are not typical of a diagnostic review; this is because the typical approach of evaluating diagnostic test accuracy against a reference standard (using sensitivity and specificity versus pathology, for example) would not be appropriate for a small metastasis; a scan can identify a real tumour which either moves or disappears before it is biopsied, and in these circumstances a negative biopsy result would not represent the gold standard; the purpose of including a list of clinical outcomes is to examine how the outcomes vary with the number of tumours detected, thus providing indirect evidence of the accuracy of the index test

Searching for evidence

Clinical search literature

Systematic literature searches were undertaken to identify all published clinical evidence relevant to the review questions.

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. All searches were conducted in MEDLINE, Embase and The Cochrane Library, with some additional database searching in AMED, PsycINFO and CINAHL for certain topic areas.

For questions where the initial search was conducted earlier than September 2017, re-run searches were carried out during September 2017.

Any studies added to the databases after the date of the last search (even those published prior to this date) were not included unless specifically stated in the text.

Search strategies were quality assured by cross-checking reference lists of highly relevant papers, analysing search strategies in other systematic reviews and asking the group members to highlight any additional studies. The questions, the study types applied, the databases searched and the years covered can be found in Appendix F in each Evidence Report.

Searching for grey literature or unpublished literature was not undertaken. Searches for electronic, ahead-of-print publications were not routinely undertaken unless requested by the committee. In this case ahead-of-print publications were sought only for review question 2c on the management of the initial diagnosis of high-grade glioma because the committee new of a large trial expected to report in early 2018. Unfortunately this paper was not possible to retrieve as the data were still being analysed by the trial team when the draft guideline for consultation was prepared.

During the scoping stage, a search was conducted for guidelines and reports on websites of organisations relevant to the topic. All references suggested by stakeholders at the scoping consultation were considered.

Health economics search literature

A global search of economic evidence was undertaken in April 2016 and re-run in September 2017. The following databases were searched:

- MEDLINE (Medical Literature Analysis and Retrieval System Online , Ovid)
- EMBASE (Excerpta Medica Database, Ovid)

- HTA database (Health Technology Appraisal database, NIHR)
- NHS EED (NHS Economic Evaluations Database, NHS).

Further to the database searches, the committee was contacted with a request for details of relevant published and unpublished studies of which they may have had knowledge; reference lists of key identified studies were also reviewed for any potentially relevant studies. Finally, the NICE website was searched for any recently published guidance relating to primary brain tumours that had not been already identified via the database searches.

The search strategy for existing economic evaluations combined terms capturing the target condition (primary brain tumours) and, for searches undertaken in MEDLINE and EMBASE, terms to capture economic evaluations. No restrictions on language or setting were applied to any of the searches, but a standard exclusions filter was applied (letters, animals, etc.). Full details of the search strategies are presented in Appendix B of each Evidence Report.

Call for evidence

No call for evidence was made.

Reviewing research evidence

Type of studies and inclusion/exclusion criteria

The evidence was reviewed following the steps shown schematically in Figure 1.

- Potentially relevant studies were identified for each review question from the relevant search results by reviewing titles and abstracts. Full papers were then obtained.
- Full papers were reviewed against pre-specified inclusion and exclusion criteria to identify studies that addressed the review question in the appropriate population, as outlined in the review protocols.
- Relevant studies were critically appraised using the appropriate checklist as specified in the NICE guidelines manual.
- Key information was extracted on the study's methods, according to the factors specified in the protocols and results. These were presented in summary tables (in each review).
- Summaries of evidence were generated by outcome (included in the relevant review chapters) and were presented to the committee as follows.
    - Randomised studies: meta-analysis was carried out where appropriate and results were reported in GRADE profiles (for intervention reviews).
    - Observational studies: if data were not suitable for pooling the effect estimate for each study with the associated confidence intervals was listed in the GRADE profiles. Data were presented as a range of values in GRADE profiles only when no confidence intervals were reported.
    - Diagnostic studies: data were presented as measures of diagnostic test accuracy (sensitivity and specificity) and were presented in modified GRADE profiles.

- Qualitative studies: each study was summarised by theme and themes were then presented in summary tables with quality ratings based on the study checklists.

Systematic reviews (SRs) with meta-analyses were considered the highest quality evidence to be selected for inclusion.

For most intervention reviews in this guideline, parallel randomised controlled trials (RCTs) were prioritised because they are considered the most robust type of study design that could produce an unbiased estimate of the intervention effects.

For diagnostic reviews, cross-sectional, retrospective or prospective observational studies were considered for inclusion. For prognostic reviews, prospective and retrospective cohort studies were included. Case-control studies were not considered for inclusion.

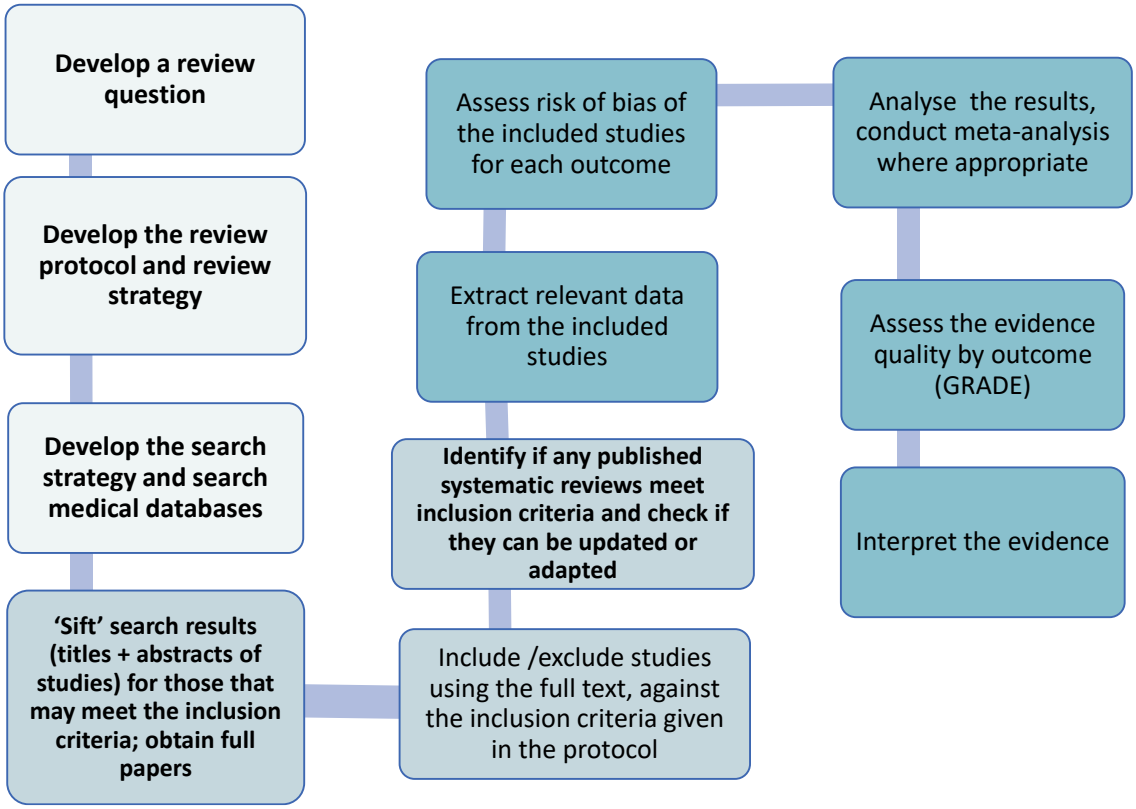
In the qualitative reviews, studies using focus groups, or structured or semi-structured interviews were considered for inclusion. Survey data or other types of questionnaires were only included if they provided analysis from open-ended questions, but not if they reported descriptive quantitative data only.

Where data from observational studies were included the results for each outcome were presented separately for each study and meta-analysis was not conducted.

For quality assurance of study identification, either whole study selections or a sample of the study selection results were double checked by a second reviewer (or a committee member with relevant expertise if the complexity of studies was high). Where this was undertaken, at least 10% of the studies were checked in this way. For details on when this was undertaken, see Appendix A of the relevant Evidence Report.

A sample of evidence tables was double extracted. All drafts of reviews were checked by a second reviewer. Any discrepancies were resolved by discussion between the reviewers.

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



Methods of combining evidence

Data synthesis for intervention studies

It was planned to conduct meta-analyses where possible, to combine the results of studies for each review question using Cochrane Review Manager (RevMan5) software.

Fixed-effect (Mantel–Haenszel) techniques were used to calculate risk ratios (relative risks (RRs) for binary outcomes, such as rate of adverse events or rate of people with symptom improvements (Mantel–Haenszel 1959).

For continuous outcomes, measures of central tendency (mean) and variation (standard deviation, SD) are required for meta-analysis. Data for continuous outcomes (such as a score on a test of neurological function) were analysed using an inverse variance method for pooling weighted mean differences. A generic inverse variance option in RevMan5 was used where any studies reported solely the summary statistics and 95% confidence interval (95% CI) or standard error; this included any hazard ratios reported. When the only evidence was based on studies summarising results by presenting medians (and interquartile ranges) or only p values were given, this information was assessed in terms of the study's sample size

and was included in the GRADE tables without calculating the relative or absolute effects. Consequently, aspects of quality assessment, such as imprecision of effect, could not be assessed for evidence of this type. However, the limited reporting of this outcome was classified as a risk of bias in study limitations.

Stratified analyses were predefined for some review questions at the protocol stage when the committee identified that strata were different in terms of biological and clinical characteristics and the interventions were expected to have a different effect.

Statistical heterogeneity was assessed by visually examining the forest plots and by considering the chi-squared test for significance at $p<0.1$ or an I-squared inconsistency statistic (with an I-squared value of more than 50% indicating high heterogeneity). Where considerable heterogeneity was present, predefined subgroup analyses were performed.

Assessments of potential differences in effect between subgroups were based on the chi-squared tests for heterogeneity statistics between subgroups. If no sensitivity analysis was found to resolve statistical heterogeneity (i.e., bring I^2 below 50%) , then a random-effects (DerSimonian and Laird) model was employed to provide a more conservative estimate of the effect (DerSimonian and Laird, 1986).

Data synthesis for diagnostic test accuracy reviews

Data and outcomes

There are a number of diagnostic test accuracy measures. Sensitivity and specificity were the main measures used as outcomes for diagnostic reviews in this guideline.

Sensitivity and specificity are measures of the ability of a test to correctly classify a person as having a disorder or not having a disorder. When sensitivity is high, a negative test result rules out the target disorder. When specificity is high, a positive test result rules in the target disorder – researchers have created the mnemonic SpPin/SnNout for this (Sackett 1992). An ideal test would be both highly sensitive and highly specific, but this is frequently not possible and typically there is a trade-off.

The area under the curve (AUC) of the receiver operating characteristic (ROC) shows true positive rate (sensitivity) as a function of false positive rate (1 minus specificity).

Data synthesis

If data are identified from more than 1 study for each of the target index tests, diagnostic paired sensitivity-specificity forest plots would have been produced for each diagnostic test using RevMan5 by extracting the 2×2 tables (the number of true positives, false positives, true negatives and false negatives). However, the evidence identified for any of the target index tests consisted only of data from single studies, so paired sensitivity and specificity estimates were simply presented in modified GRADE tables.

Data synthesis for qualitative reviews

Where possible, a meta-synthesis was conducted to combine qualitative study results. The main aim of the synthesis of qualitative data was to produce a

description of topics that may influence the experience of follow up for a person with a brain tumour, people important to them and healthcare professionals involved in their care, rather than building new theories or reconceptualising the topic under review. Whenever studies identified a qualitative theme, this was extracted and the main characteristics were summarised. When all themes were extracted from studies, common concepts were categorised and tabulated. This included information on how many studies had contributed to an identified overarching theme. In qualitative synthesis, a theme being reported by different studies more often than other themes does not necessarily mean that it would be more important than those other themes. The aim of qualitative research is to identify new perspectives on a particular topic. Study type and population in qualitative research can differ widely, meaning that themes identified by just 1 or a few studies can provide important new information for a given topic.

The most relevant evidence in this respect would originate from studies set in the target context of the UK NHS setting. Themes from individual studies were then integrated into a wider context and, when possible, overarching categories of themes with subthemes were identified. Themes were derived from data presented in individual studies based directly on quotes from interviewees. When themes were extracted, the names attached to the themes in the studies in which they were derived were used in the guideline. The names of overarching themes, however, were added by the guideline reviewers. Data saturation was examined on a study level and based on the definitions and assessments made by the individual study authors. All relevant studies were included in the qualitative reviews with no studies excluded based on data saturation at review level (i.e. study inclusion did not stop even if no new themes were identified in the subsequently added studies).

Appraising the quality of evidence

Intervention studies

GRADE methodology (The Grading of Recommendations Assessment, Development and Evaluation)

For intervention reviews, the evidence for outcomes from the included RCTs was evaluated and presented using GRADE, which was developed by the international GRADE working group.

The software developed by the GRADE working group (GRADEpro) was used to assess the quality of each outcome, taking into account individual study quality factors and the meta-analysis results. The clinical/economic evidence profile tables include details of the quality assessment and pooled outcome data, where appropriate, an absolute measure of intervention effect and the summary of quality of evidence for that outcome. In this table, the columns for intervention and control indicate summary measures of effect and measures of dispersion (such as mean and SD or median and range) for continuous outcomes and frequency of events (n/N; the sum across studies of the number of patients with events divided by sum of the number of completers) for binary outcomes. Reporting or publication bias was only taken into consideration in the quality assessment and included in the clinical evidence profile tables if it was apparent.

The selection of outcomes for each review question was decided when each review protocol was discussed with the guideline committee.

The evidence for each outcome in the intervention reviews was examined separately for the quality elements listed and defined in Table 2. Each element was graded using the quality levels listed in Table 3.

The main criteria considered in the rating of these elements are discussed below. Footnotes were used to describe reasons for grading a quality element as having serious or very serious limitations. The ratings for each component were summed to obtain an overall assessment for each outcome (Table 4).

Table 2: Description of quality elements in GRADE for intervention reviews

Quality element	Description
Risk of bias (study limitations)	Limitations in the study design and implementation may bias the estimates of the treatment effect. High risk of bias for the majority of the evidence decreases confidence in the estimate of the effect.
Inconsistency	Inconsistency refers to an unexplained heterogeneity of results or findings.
Indirectness	Indirectness refers to differences in study population, intervention, comparator and outcomes between the available evidence and the review question, or recommendation made, such that the effect estimate is changed. This is also related to applicability or generalisability of findings.
Imprecision	Results are imprecise when studies include relatively few patients and few events and thus have wide confidence intervals around the estimate of the effect. Imprecision results if the confidence interval includes the clinically important threshold.
Publication bias	Publication bias is a systematic underestimate or an overestimate of the underlying beneficial or harmful effect due to the selective publication of studies.

Table 3: Levels of quality elements in GRADE

Levels of quality elements in GRADE	Description
None/no serious	There are no serious issues with the evidence.
Serious	The issues are serious enough to downgrade the outcome evidence by 1 level.
Very serious	The issues are serious enough to downgrade the outcome evidence by 2 levels.

Table 4: Levels of overall quality of outcome evidence in GRADE

Overall quality of outcome evidence in GRADE	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.

Overall quality of outcome evidence in GRADE	Description
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.

Assessing risk of bias in intervention reviews

Bias is a systematic error, or a consistent deviation from the truth in the results. When a risk of bias is present the true effect can be either under- or over-estimated.

Risk of bias in intervention studies was assessed using the Cochrane Risk of Bias Tool (see Appendix H in the NICE guidelines manual 2014).

It should be noted that a study with a poor methodological design does not automatically imply high risk of bias; the bias is considered individually for each outcome and it is assessed whether this poor design will impact on the estimation of the intervention effect.

Assessing inconsistency in intervention reviews

Inconsistency refers to unexplained heterogeneity of results of meta-analysis. When estimates of the treatment effect vary widely across studies (that is, there is heterogeneity or variability in results), this suggests true differences in underlying effects. Inconsistency is, thus, only applicable when statistical meta-analysis is conducted (that is, results from different studies are pooled). However, 'no inconsistency' is nevertheless used to describe this quality assessment in the GRADE profiles for outcomes from single studies as this is the default option in the GRADEpro software used.

Heterogeneity was assessed by calculating the I-squared statistic for the meta-analysis. I-squared values of more than 50% and 80% were considered to indicate high and very high heterogeneity, respectively. When high or very high heterogeneity was observed, possible reasons for it were explored and subgroup analyses were performed as pre-specified in the review protocol.

When no plausible explanation for the heterogeneity could be found, the quality of the evidence was downgraded in GRADE by 1 (I-squared > 50%) or 2 (I-squared > 80%) levels for the domain of inconsistency, depending on the extent of heterogeneity in the results.

Assessing indirectness in intervention reviews

Directness refers to the extent to which the populations, intervention, comparisons and outcome measures are similar 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.

Assessing imprecision and clinical significance in intervention reviews

Imprecision in guidelines concerns whether the uncertainty (CI) around the effect estimate means that it is not clear whether there is a clinically important difference between interventions or not (that is, whether the evidence would clearly support one recommendation or appear to be consistent with several different types of recommendations). Therefore, imprecision differs from the other aspects of evidence quality because it is not really concerned with whether the point estimate is accurate or correct (has internal or external validity). Instead, it is concerned with the uncertainty about what the point estimate actually is. This uncertainty is reflected in the width of the CI.

The 95% CI is defined as the range of values within which the population value will fall on 95% of repeated samples, were this procedure to be repeated. The larger the trial, the smaller the 95% CI and the more certain the effect estimate.

Imprecision in the evidence reviews was assessed by considering whether the width of the 95% CI of the effect estimate was relevant to decision-making, taking each outcome in isolation. This is explained in Figure 2, which considers a positive outcome for the comparison of treatment A versus treatment B. Three decision-making zones can be identified, bounded by the thresholds for clinical importance (minimally important difference, MID) for benefit and for harm. The MID for harm for a positive outcome means the threshold at which drug A is less effective than drug B by an amount that is clinically important to patients (favours B).

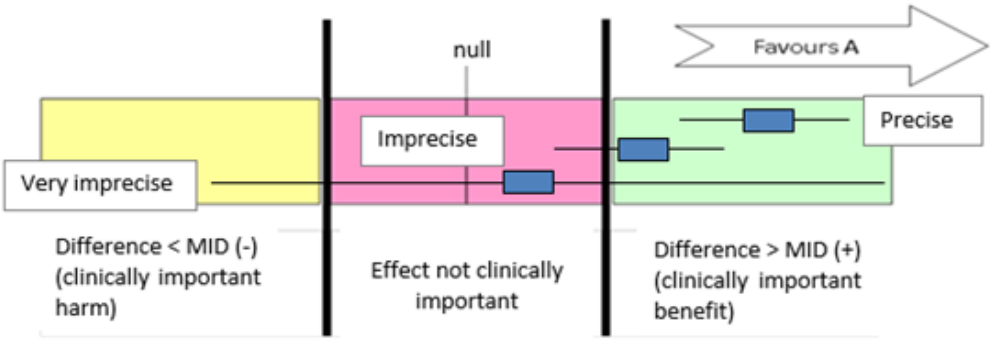
When the CI of the effect estimate is wholly contained in 1 of the 3 zones (for example, clinically important benefit), we are not uncertain about the size and direction of effect (whether there is a clinically important benefit, or the effect is not clinically important, or there is a clinically important harm), so there is no imprecision.

When a wide CI lies partly in each of 2 zones, it is uncertain in which zone the true value of effect estimate lies and therefore there is uncertainty over which decision to make (based on this outcome alone). The CI is consistent with 2 possible decisions and so this is considered to be imprecise in the GRADE analysis and the evidence is downgraded by 1 level (‘serious imprecision’).

If the CI of the effect estimate crosses into 3 zones, this is considered to be very imprecise evidence because the CI is consistent with 3 possible clinical decisions and there is therefore a considerable lack of confidence in the results. The evidence is therefore downgraded by 2 levels in the GRADE analysis (‘very serious imprecision’).

Implicitly, assessing whether the CI is in, or partially in, a clinically important zone, requires an MID to be defined or to say whether they would make different decisions for the 2 confidence limits.

Figure 2: Illustration of precise, imprecise and very imprecise evidence based on the confidence interval of outcomes in forest plots



Minimally important differences

The literature was searched for established MIDs for the selected outcomes in the evidence reviews, such as cognitive function or quality of life. In addition, the committee members were asked whether they were aware of any acceptable MIDs in the clinical community.

As no published or acceptable MIDs were identified, the committee decided that it was clinically acceptable to use the GRADE default MID to assess imprecision. For binary outcomes clinically important thresholds for a RR of 0.8 and 1.25 respectively were used (due to the statistical distribution of this measure this means that this is not a symmetrical interval). In the absence of published and default GRADE MIDs for HRs, and on the basis of consultation with the NGA statistician, the same MID was used to assess imprecision of HRs. Thus, this default MID was used for all the binary outcomes in the intervention evidence reviews as a starting point and decisions on clinical importance were then considered based on the absolute risk difference. For continuous outcomes GRADE default MIDs were half of the median SD of the control groups across trials (when more than one trial reported the same outcome).

Diagnostic studies

The GRADE toolbox is designed for RCTs and observational studies, but we adapted the quality assessment elements and outcome presentation for diagnostic test accuracy reviews. For example, the GRADE clinical evidence tables were modified to include the most appropriate measures of diagnostic accuracy (sensitivity, specificity, and likelihood ratios).

The evidence for each outcome in the diagnostic test accuracy reviews was examined separately for the quality elements listed and defined in Table 5. Each element was graded using the quality levels listed in Table 3.

The main criteria considered in the rating of these elements are discussed below. Footnotes were used to describe reasons for grading a quality element as having serious or very serious limitations. The ratings for each component were summed to obtain an overall assessment for each outcome (Table 4).

Table 5: Description of the elements in GRADE and how they are used to assess the quality for diagnostic accuracy reviews

Quality element	Description
Risk of bias ('Study limitations')	Limitations in the study design and implementation may bias the estimates of the diagnostic accuracy. High risk of bias for the majority of the evidence decreases confidence in the estimate of the effect. Diagnostic accuracy studies are not usually randomised and therefore would not be downgraded for study design from the outset and start as high level evidence.
Inconsistency	Inconsistency refers to an unexplained heterogeneity of test accuracy measures, for example sensitivity or specificity, between studies.
Indirectness	Indirectness refers to differences in study population, index tests, reference standards and outcomes between the available evidence and the review question.
Imprecision	Results are considered imprecise when studies include relatively few patients and the probability to be diagnosed correctly in this group is low. Imprecision results if the confidence interval includes the clinically important threshold.

Assessing risk of bias and indirectness in diagnostic test accuracy reviews

Risk of bias in diagnostic test accuracy studies was assessed using the Quality Assessment of Diagnostic Accuracy Studies version 2 (QUADAS-2) checklist (see Appendix H in the NICE guidelines manual 2014).

Risk of bias and applicability in primary diagnostic accuracy studies in QUADAS-2 consists of 4 domains (risk of bias and applicability are assessed for the first 3 domains, with only risk of bias assessed for the fourth domain:

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

More details about the quality assessment of diagnostic studies are shown in Table 6.

Table 6: Summary of QUADAS-2

Domain	Patient Selection	Index text	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 2x2 table: Describe the time interval and any interventions

Domain	Patient Selection	Index text	Reference standard	Flow and timing
				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?	

Assessing inconsistency in diagnostic test accuracy reviews

Inconsistency refers to unexplained heterogeneity of the results in meta-analysis. When estimates of diagnostic accuracy parameters vary widely across studies (that is, there is heterogeneity or variability in results), this suggests true differences in underlying effects. Inconsistency is, thus, only applicable when statistical meta-analysis is conducted (that is, results from different studies are pooled). However, 'no inconsistency' is nevertheless used to describe this quality assessment in the GRADE profiles for outcomes from single studies.

For the diagnostic test accuracy reviews, no meta-analyses were conducted, thus no inconsistency judgements were made.

Assessing indirectness in diagnostic test accuracy reviews

Indirectness in diagnostic test accuracy studies was assessed using the QUADAS-2 checklist by assessing the applicability of the studies in relation to the review question in the following domains (see Table 6):

- patient selection
- index test
- reference standard.

Assessing imprecision in diagnostic test accuracy reviews

In evaluating diagnostic test accuracy measures, it was first decided by the committee that sensitivity should be given the most weight in the decision-making process because the committee agreed that the consequences of a false negative test (i.e., a missed diagnosis) were more harmful to patients than the consequences of a false positive test (i.e., a false diagnosis). Imprecision was, therefore, rated on this statistical measure using the following criterion:

- if the difference between 95% CI confidence limits for sensitivity was ≥ 25% (that is, the CI was very wide) then the quality of the evidence was downgraded by 1 level.

Qualitative reviews

For qualitative evidence, quality was assessed using a checklist for qualitative studies (NICE 2015). This was based on the Critical Appraisal Skills Programme (CASP) checklist for qualitative studies (Table 7). The quality rating for risk of bias (low, high and unclear) was derived by assessing the risk of bias across 6 domains.

The evidence was then assessed by theme using the ratings on the CASP checklist across studies taking into account any identified limitations as described in Table 7 and labelled as low (more than one study limitation identified), moderate (one study limitation identified) or high quality (no study limitations identified).

Table 7: Summary of CASP tool for qualitative studies

Risk of bias	Explanation
Aim and appropriateness of qualitative evidence.	This refers to an assessment of whether the aims and relevance of the study were clearly described and whether qualitative research methods were appropriate for investigating the research question.
Rigour in study design or validity of theoretical approach	This domain assesses whether the study approach has been clearly described and is based on a theoretical framework (for example ethnography or grounded theory). This does not necessarily mean that the framework has to be explicitly stated, but that at least a detailed description is provided which makes it transparent and reproducible.
Sample selection	The background, the procedure and reasons for the chosen method of selecting participants should be stated. It should also be assessed whether there was a relationship between the researcher and the informant and if so, how this may have influenced the findings that were described.

Risk of bias	Explanation
Data collection	Consideration was given to how well the method of data collection (in-depth interviews, semi-structured interviews, focus groups or observations) was described, whether details were provided and how the data were collected (who conducted the interviews, how long did they last and where did they take place).
Data analysis	For this criterion it is assessed whether sufficient detail is provided about the analytical process and whether it is in accordance with the theoretical approach. For instance, if a thematic analysis was used, it is assessed whether there was a clear description of how the theme was arrived at. Data saturation is also part of this section. This refers to whether a theoretical point of theme saturation was achieved at which point no further citations or observations would provide more insight or suggest a different interpretation of this theme. This could be explicitly stated, or it may be clear from the citations presented that it may have been possible to find more themes.
Results	In relation to this section the reasoning about the results are important, for instance whether a theoretical proposal or framework is provided rather than being restricted to citations / presentation of data.

Evidence statements

Evidence statements are summary statements that are presented after the GRADE profiles highlighting the key features of the clinical evidence presented. The wording of the evidence statements reflects the certainty or uncertainty in the estimate of effect. The evidence statements are presented by outcome or theme and encompass the following key features of the evidence:

- the quality of the evidence (GRADE rating)
- the number of studies and/or the number of participants for a particular outcome (or theme in the case of qualitative evidence)
- a brief description of the participants
- the clinical significance of the effect and an indication of its direction (for example, if a treatment is clinically significant (beneficial or harmful) compared with another, or whether there is no clinically significant difference between the tested treatments).

Reviewing economic evidence

Inclusion and exclusion of economic studies

The titles and abstracts of papers identified through the searches were independently assessed for inclusion using predefined eligibility criteria defined in Table 8.

Table 8: Inclusion and exclusion criteria for the systematic reviews of economic evaluations

Inclusion criteria
Economic evaluations that compare costs and health consequences of interventions (i.e. true cost-effectiveness analyses)
Population, interventions, comparators and outcomes match those specified in the PICO

Quality of life based outcomes were used as the measure of effectiveness in at least one of the analyses presented
Incremental results reported or enough information for incremental results to be derived
Conducted from the perspective of a healthcare system in an OECD country
Exclusion criteria
Conference abstracts with insufficient methodological details for quality assessment
Non-English language papers

OECD Organisation for Economic Co-operation and Development; PICO Population, Intervention, Comparison, and Outcome.

Once the screening of titles and abstracts was complete, full versions of the selected papers were acquired for assessment. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) for the search on economic evaluations is presented in Supplementary Material D.

The quality of evidence was assessed using the economic evaluations checklist as specified in the NICE guidelines manual (NICE 2014). Quality assessments of included studies and data extraction tables are provided in Appendix F of the relevant Evidence Report and Supplementary Material D respectively. The excluded economic studies list is presented in Appendix K of the respective Evidence Report.

Appraising the quality of economic evidence

The quality of economic evaluations in this guideline were appraised using the methodology checklist reported in the NICE Guideline Manual 2014, Appendix H for all studies which met the inclusion criteria.

Health economic modelling

The aims of the health economic input to the guideline were to inform the guideline committee of potential economic issues related to primary brain tumours and brain metastases in adults to ensure that recommendations represented a cost-effective use of healthcare resources. Health economic evaluations aim to integrate data on healthcare benefits (ideally in terms of quality-adjusted life-years, QALYs) with the costs of different care options. In addition, the health economic input aimed to identify areas of high resource impact; recommendations which – while nevertheless cost-effective – might have a large impact on Clinical Commissioning Group or Trust finances and so need special attention.

The committee prioritised two economic models on the treatment of a single metastasis and resection of high-grade glioma for which they thought economic considerations would be particularly important when formulating recommendations.

The methods and results of the de novo economic analyses are reported in Appendix J of Evidence Report A and Evidence Report C, respectively. When new economic analysis was not prioritised, the committee made a qualitative judgement regarding cost effectiveness by considering expected differences in resource and cost use between options, alongside clinical effectiveness evidence identified from the clinical evidence review.

Cost effectiveness criteria

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

- 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, or
- the intervention provided clinically significant benefits at an acceptable additional cost when compared with the next best strategy.

The committee's considerations of cost-effectiveness are discussed explicitly under the 'Cost Effectiveness and Resource Use' headings of the relevant sections.

Developing recommendations

Guideline recommendations

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

The main considerations specific to each recommendation are outlined under the 'Recommendations and link to evidence' headings within each Evidence Report.

For further details please refer to the NICE guidelines manual (NICE 2014).

Research recommendations

When areas were identified for which good evidence was lacking, the committee considered making recommendations for future research. For further details please refer to the NICE guidelines manual (NICE 2014).

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 at publication. For further details please refer to the NICE guidelines manual (NICE 2014).

Updating the guideline

Following publication, and in accordance with the NICE guidelines manual, NICE will undertake a review of whether the evidence base has progressed significantly to alter the guideline recommendations and warrant an update. For further details please refer to the NICE guidelines manual (NICE 2014).

Funding

The NGA was commissioned by NICE to develop this guideline.

References

Bernstein 1992

Bernstein SJ, Laouri M, Hilborne LH, Leape LL, Kahan JP, Park R, et al. Coronary angiography: a literature review and ratings of appropriateness and necessity. Santa Monica, USA: RAND; 1992.

Debray 2017

Debray, T., Damen, J., Snell, K., Ensor, J., Hooft, L., Reitsma, J. et al. A guide to systematic review and meta-analysis of prediction model performance BMJ, 356, i6460, 2017.

DerSimonian and Laird 2015

DerSimonian R, Laird N. Meta-analysis in clinical trials revisited. Contemporary clinical trials. 2015 Nov 30; 45:139-45.

Murphy 1998

Murphy MK, Black NA, Lamping D, McKee C, Sanderson C, Askham J, et al. Consensus development methods, and their use in clinical guideline development. Health Technology Assessment. 1998; 2:i-88.

NICE 2014

National Institute for Health and Care Excellence (NICE), NICE guidelines manual, 2014 [online, accessed 21st November 2017]

NICE 2008

National Institute for Health and Care Excellence (NICE), Social Value Judgements: Principles of the development of NICE guidance [online, accessed 21st November 2017]

Santesso 2016

Santesso, N., Carrasco-Labra, A., Langendam, M., Brignardello-Petersen, R., Mustafa, R.A., Heus, P., Lasserson, T., Opiyo, N., Kunnamo, I., Sinclair, D. and Garner, P., Improving GRADE evidence tables part 3: detailed guidance for explanatory footnotes supports creating and understanding GRADE certainty in the evidence judgments, Journal of clinical epidemiology, 74, 28-39, 2016.

Schünemann 2013

Schünemann, H., Brożek, J, Guyatt, G., Oxman, A., (eds.). Handbook for grading the quality of evidence and the strength of recommendations using the GRADE approach (updated October 2013). GRADE Working Group, 2013.

Wells 2008

Wells GA, Shea B, O'Connell D, Peterson J, Welch V, Losos M, Tugwell P. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses.