National Institute for Health and Care Excellence

Final

Colorectal cancer (update)

Supplement 2: Methods

NICE guideline NG151

Methods

January 2020

Final

Developed by the National Guideline Alliance part of 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 All NICE guidance is subject to regular review and may be updated or withdrawn.

ISBN: 978-1-4731-3657-1

Contents

Development of the guideline

Remit

The National Institute for Health and Care Excellence (NICE) commissioned the National Guideline Alliance (NGA) to develop a guideline to update and replace the NICE guideline on colorectal cancer: diagnosis and management (CG131) and the NICE guideline on improving outcomes in colorectal cancer (CSG5).

Declarations of interest

Committee members' and developers' declarations of interest were recorded according to NICE's 2014 conflicts of interest policy until 31st March 2018, and thereafter in accordance with NICE's 2018 conflicts of interest policy.

What this guideline covers

Groups that are covered

- Adults (18 years and older) with newly diagnosed adenocarcinoma of the colon
- Adults with newly diagnosed adenocarcinoma of the rectum
- Adults with relapsed adenocarcinoma of the colon
- Adults with relapsed adenocarcinoma of the rectum
- Adults with clinical or genetic evidence of Lynch syndrome (hereditary nonpolyposis colorectal cancer, HNPCC)

Clinical areas that are covered

The guideline covers the following clinical issues:

- Prevention of colorectal cancer
  - Role of aspirin in the prevention of colorectal cancer in adults with clinical or genetic evidence of Lynch syndrome (hereditary nonpolyposis colorectal cancer)
- Molecular biomarkers
  - Use of molecular biomarkers to guide chemotherapy choice
- Management of local disease
  - Rectal cancer
  - Colon cancer
  - Colonic stents for obstructing colon cancer
- Management of metastatic disease
  - Presenting with stage IV colorectal cancer
  - Methods for treating metastasis
- Ongoing care and support
  - Follow-up after apparently curative resection
  - Management of post treatment sequelae
  - Information about managing bowel function

- Service delivery
  - Surgical volumes and rectal cancer surgery

For further details please refer to the scope on the NICE website.

What this guideline does not cover

Groups that are not covered

The guideline does not cover the following groups:

- People with anal cancer
- Children and young people aged under 18 years
- People with primary or secondary lymphoma of the colon and rectum
- People with pure small cell carcinoma, or other pure neuroendocrine carcinomas, of the colon and rectum
- People with neuroendocrine tumours of the colon and rectum
- People with gastrointestinal stromal tumours (GIST) or sarcoma of the colon and rectum
- People with squamous cells carcinoma of the rectum
- People with appendiceal neoplasms

Clinical areas that are not covered

This guideline does not cover the following areas:

- Population screening
- Colonoscopic surveillance of high-risk groups, including people with a family history of colorectal cancer and people with inflammatory bowel disease
- Management of anal cancer

The following areas covered by CG131 were not updated and will be removed from the guideline as there is no longer variation in practice:

- Diagnostic investigations
- Staging of colorectal cancer
- Imaging of hepatic metastases
- Imaging of extra-hepatic metastases

The following areas from CSG5 will not be updated either because they are already covered within scope of update of CG131 or other NICE guidelines or because they are no longer relevant to this guideline:

- Patient centred care
- Access to appropriate services
- Multidisciplinary teams
- Diagnosis
- Surgery and histopathology
- Radiotherapy in primary disease
- Adjuvant chemotherapy

- Anal cancer
- Follow-up
- Recurrent and advances disease
- Palliative care

For further details please refer to the scope on the NICE website.

Methods

Introduction

This section summarises methods used to identify and review the evidence, to consider cost effectiveness, and to develop guideline recommendations. This guideline was developed in accordance with methods described in Developing NICE guidelines: the manual (NICE 2014).

Developing the review questions and outcomes

The review questions considered in this guideline were based on the key areas identified in the guideline scope .They were drafted by the NGA technical team, and refined and validated by the guideline committee.

The review questions were based on the following frameworks:

- intervention reviews – using population, intervention, comparison and outcome (PICO)
- prognostic reviews – using population, presence or absence of a prognostic, risk or predictive factor and outcome (PPO)
- qualitative review – using population, phenomenon of interest and context (PICo).

These frameworks guided the development of review protocols, the literature searching process, and critical appraisal and synthesis of evidence. They also facilitated development of recommendations by the committee.

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

The review questions and evidence review reports corresponding to each question (or group of questions) are summarised in Table 1.

Table 1: Summary of review questions and index to evidence reports

Evidence report	Review question	Type of review
A. Prevention of colorectal cancer		
A1. Effectiveness of aspirin in the prevention of colorectal cancer in people with Lynch syndrome	How effective is aspirin in the prevention of colorectal cancer in adults with Lynch syndrome (hereditary nonpolyposis colorectal cancer)?	Intervention
B. Molecular biomarkers		
B1. Use of molecular biomarkers to guide systemic therapy	Which predictive biomarkers should be used in the systemic management of colorectal cancer patients?	Predictive/prognostic
C. Management of local disease		
C1. Treatment for early rectal cancer	What is the most effective treatment for early rectal cancer?	Intervention

Evidence report	Review question	Type of review
C2. Preoperative radiotherapy and chemoradiotherapy for rectal cancer	What is the effectiveness of preoperative radiotherapy or chemo radiotherapy for rectal cancer?	Intervention
C3. Optimal surgical technique for rectal cancer	What is the optimal surgical technique for rectal cancer?	Intervention
C4. Deferral of surgery in people having neoadjuvant therapy for rectal cancer	Which people having neoadjuvant chemotherapy or chemoradiotherapy for rectal cancer do not need surgery?	Prognostic
C5. Effectiveness of exenterative surgery for locally advanced or recurrent rectal cancer	What is the effectiveness of exenterative surgery for locally advanced or recurrent rectal cancer?	Intervention
C6. Endoscopic resection alone for early colon cancer	Which people with early colon cancer can be treated with endoscopic resection alone?	Intervention
C7. Preoperative chemotherapy for non-metastatic colon cancer	Which people with non-metastatic colon cancer would benefit from preoperative chemotherapy?	Intervention
C8. Optimal duration of adjuvant chemotherapy for colorectal cancer	What is the optimal duration of adjuvant chemotherapy for colorectal cancer?	Intervention
C9. Effectiveness of stenting for acute large bowel obstruction	What is the effectiveness of stenting compared with emergency surgery for suspected colorectal cancer causing acute large bowel obstruction?	Intervention
D. Management of metastatic disease		
D1. Surgery for asymptomatic primary tumour	Does surgery for the asymptomatic primary tumour improve outcomes for people with metastatic colorectal cancer, which cannot be treated with curative intent?	Intervention
D2a. Treatment for metastatic colorectal cancer in the liver amenable to treatment with curative intent	What is the optimal combination and sequence of treatments in patients presenting with metastatic colorectal cancer in the liver amenable to treatment with curative intent?	Intervention
D2b. Treatment for metastatic colorectal cancer in the liver not amenable to	What is the optimal combination and sequence of treatments in patients presenting with metastatic colorectal cancer in the liver not amenable to treatment with curative intent?	Intervention

Evidence report	Review question	Type of review
treatment with curative intent		
D3. Treatment for metastatic colorectal cancer in the lung amenable to local treatment	What is the optimal combination and sequence of treatments in patients presenting with metastatic colorectal cancer in the lung amenable to local treatment?	Intervention
D4. Local and systemic treatments for metastatic colorectal cancer isolated in the peritoneum	What is the optimal combination and sequence of local and systemic treatments in patients presenting with metastatic colorectal cancer isolated in the peritoneum?	Intervention
E. Ongoing care and support		
E1. Follow-up to detect recurrence after treatment for non-metastatic colorectal cancer	What are the optimal methods and frequencies of follow-up to detect recurrence after potentially curative surgical treatment for non-metastatic colorectal cancer?	Intervention
E2. Optimal management of low anterior resection syndrome	What is the optimal management of low anterior resection syndrome?	Intervention
E3. Information needs of people prior, during and after treatment for colorectal cancer	What are the information needs of people prior, during and after treatment for colorectal cancer?	Qualitative
F. Service delivery		
F1. Surgical volumes and outcomes for rectal cancer	Is there a relationship between surgical volumes and outcomes in the treatment of rectal cancer (primary and recurrent disease)?	Predictive/prognostic

Additional information related to development of the guideline is contained in:

- Supplement 1: Methods (this document)
- Supplement 2: Health economics
- Supplement 3: Glossary
- Supplement 4: NGA technical team list.

Searching for evidence

Clinical literature search

Systematic literature searches were undertaken to identify published clinical evidence relevant to each review question. Combined searches were more than one review question were conducted where appropriate. A combined search was done for evidence reviews C1, C2 and C3; and for evidence reviews D2a and D2b.

Databases were searched using medical subject headings, free-text terms and study type filters where appropriate. Where possible, searches were restricted to retrieve articles published in English. All searches were conducted in the following databases: Medline, Medline-in-Process, Cochrane Central Register of Controlled Trials (CCTR), Cochrane Database of Systematic Reviews (CDSR), Database of Abstracts of Reviews of Effects (DARE), Health Technology Assessments (HTA) and Embase. For review questions related to information provision, PsycInfo, CINAHL and Web of Science were also searched. Web of Science was also used for the question about prevention of colorectal cancer.

Searches were run once for all reviews during development. Searches were updated 6 to 8 weeks in advance of the final committee meetings for the following questions:

- C1. What is the most effective treatment for early rectal cancer?
- C2. What is the effectiveness of preoperative radiotherapy or chemo radiotherapy for rectal cancer?
- C3. What is the optimal surgical technique for rectal cancer?
- C4. Which people having neoadjuvant chemotherapy or chemoradiotherapy for rectal cancer do not need surgery?
- C5. What is the effectiveness of exenterative surgery for locally advanced or recurrent rectal cancer?
- C8. What is the optimal duration of adjuvant chemotherapy for colorectal cancer?
- D1. Does surgery for the asymptomatic primary tumour improve outcomes for people with metastatic colorectal cancer, which cannot be treated with curative intent?
- D2a. What is the optimal combination and sequence of treatments in patients presenting with metastatic colorectal cancer in the liver amenable to treatment with curative intent?
- D2b. What is the optimal combination and sequence of treatments in patients presenting with metastatic colorectal cancer in the liver not amenable to treatment with curative intent?
- E1. What are the optimal methods and frequencies of follow-up to detect recurrence after potentially curative surgical treatment for non-metastatic colorectal cancer?
- E2. What is the optimal management of low anterior resection syndrome?
- F1. Is there a relationship between surgical volumes and outcomes in the treatment of rectal cancer (primary and recurrent disease)?

Literature searches were not updated for the remaining review questions because:

- the original search was done within 8 weeks of the final committee meeting
- there was already robust evidence and new evidence would not change the conclusions or the committee was confident there was no new published evidence.

Search strategies were quality assured by cross-checking reference lists of relevant articles, analysing search strategies from other systematic reviews and asking members of the committee to highlight key studies. All search strategies were also quality assured by an information scientist who was not involved in developing the primary search strategy. Details of the search strategies, including study-design filters applied and databases searched, are presented in Appendix B of each evidence report.

All publications highlighted by stakeholders at the time of the consultation on the draft scope were considered for inclusion. During the scoping phase, searches were conducted for guidelines, health technology assessments, systematic reviews, economic evaluations and reports on biomedical databases and websites of organisations relevant to the topic. Formal searching for grey literature and unpublished literature was not undertaken routinely.

Health economic literature search

A global search of economic evidence was undertaken and re-run in May 2019. The following databases were searched:

- MEDLINE (Ovid)
- EMBASE (Ovid)
- Health Technology Assessment database (HTA)
- NHS Economic Evaluation Database (NHS EED)

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 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 colorectal cancer that had not been already identified via the database searches.

The search strategy for existing economic evaluations combined terms capturing colorectal cancer and, for searches undertaken in MEDLINE and EMBASE, terms to capture economic evaluations. No restrictions on language or setting were applied to the economic evidence search, but a standard exclusions filter was applied (letters, animals, etc.). Full details of the search strategy are presented in Supplement 2: Health economics.

Reviewing research evidence

Systematic review process

The evidence was reviewed in accordance with the following approach.

- Potentially relevant articles were identified from the search results for each review question by screening titles and abstracts. Full-text copies of the articles were then obtained.
- Full-text articles were reviewed against pre-specified inclusion and exclusion criteria in the review protocol (see appendix A of each evidence report).
- Key information was extracted from each article on study methods and results, in accordance with factors specified in the review protocol. The information was presented in a summary table in the corresponding evidence report and in a more detailed evidence table (see appendix D of each evidence report).
- Included studies were critically appraised using an appropriate checklist as specified in Developing NICE guidelines: the manual (NICE 2014). Further detail on appraisal of the evidence is provided below.
- Summaries of evidence by outcome were presented in the corresponding evidence report and discussed by the committee.

All review questions were subject to dual screening and study selection through a 10% random sample of articles, except for the question on effectives of aspirin in the prevention of colorectal cancer in adults with Lynch syndrome. This question was not dual screened because the intervention, population and study design were well defined. Any discrepancies in dual screening were resolved by discussion between the first and second reviewers or by reference to a third (senior) reviewer. Internal (NGA) quality assurance processes included consideration of the outcomes of screening, study selection and data extraction and the committee reviewed the results of study selection and data extraction. The review protocol for each question specifies whether dual screening and study selection was undertaken for that particular question.

Drafts of all evidence reviews were checked by a senior reviewer.

Type of studies and inclusion/exclusion criteria

Inclusion and exclusion of studies was based on criteria specified in the corresponding review protocol.

Systematic reviews with meta-analyses were considered to be the highest quality evidence that could be selected for inclusion.

For intervention reviews, randomised controlled trials (RCTs) were prioritised for inclusion because they are considered to be the most robust type of study design that could produce an unbiased estimate of intervention effects. Where there was limited evidence from RCTs, non-randomised controlled trials were considered for inclusion.

For prognostic reviews, prospective and retrospective cohort and case–control studies and case series were considered for inclusion.

For the qualitative review, studies using focus groups, structured interviews or semi-structured interviews were considered for inclusion. Where qualitative evidence was sought, data from surveys or other types of questionnaire were considered for inclusion only if they provided data from open-ended questions, but not if they reported only quantitative data.

The committee was consulted about any uncertainty regarding inclusion or exclusion of studies. A list of excluded studies for each review question, including reasons for exclusion is presented in appendix K of the corresponding evidence report.

Narrative reviews, posters, letters, editorials, comment articles, unpublished studies and studies published in languages other than English were excluded. Conference abstracts were not considered for inclusion.

Methods of combining evidence

When planning reviews (through preparation of protocols), the following approaches for data synthesis were discussed and agreed with the committee.

Data synthesis for intervention studies

Meta-analysis to pool results from RCTs was conducted where possible using Cochrane Review Manager (RevMan5) software. Where non-randomised evidence

was used, this was meta-analysed only if results had been adjusted for pre-specified confounders.

For dichotomous outcomes, such as mortality, the Mantel–Haenszel method with a fixed effect model was used to calculate risk ratios (relative risks; RRs). For all outcomes with zero events in both arms the risk difference was presented. For outcomes with zero events in only one arm, Peto odds ratios (ORs) were calculated as this method performs well when events are rare (Bradburn 2007).

In some cases outcomes were meta-analysed for a single study arm only. This was when the outcome was only possible in one study arm, for example the technical success of stenting in trials comparing stents to palliative care. Studies were only meta-analysed if they had similar populations, interventions and outcome definitions. Proportions were first logit transformed and then meta-analysed with a random-effects model using the metafor package in R (Viechtbauer 2010).

Time-to-event outcomes, such as overall survival, were meta-analysed using the "O – E (Observed – Expected) and Variance" outcome type in RevMan5. This provides a fixed-effect meta-analysis of log hazard ratios. Corrections for zero cell counts are not needed when using this method. If O – E and Variance were not reported in a study, where possible these were calculated using methods reported by Tierney (2007). In some cases where studies had used Cox proportional hazards regression models, generic inverse variance meta-analysis was used. Both fixed-effect and random-effects analyses were available in this case.

For continuous outcomes, measures of central tendency (mean) and variation (standard deviation; SD) are required for meta-analysis. Data for continuous outcomes, such as duration of hospital stay, were meta-analysed using an inverse-variance method for pooling weighted mean differences (WMDs). Where SDs were not reported for each intervention group, the standard error (SE) of the mean difference was calculated from other reported statistics (p values or 95% confidence intervals; CIs) and then meta-analysis was conducted as described above.

When evidence was based on studies that reported descriptive data or medians with interquartile ranges or p values, this information was included in the corresponding GRADE tables (see below) without calculating relative or absolute effects. Consequently, certain aspects of quality assessment such as imprecision of the effect estimate could not be assessed for this type of evidence. The limited reporting was interpreted as representing a risk of bias when assessing study limitations.

Subgroups for stratified analyses were agreed for some review questions as part of protocol development.

When meta-analysis was undertaken, the results were presented visually using forest plots generated using RevMan5 (see appendix E of relevant evidence reports).

Data synthesis for prognostic reviews

Odds ratios (ORs) or RRs with 95% CIs reported in published studies were extracted or calculated by the NGA technical team to examine relationships between risk factors and outcomes of interest. Ideally analyses would have adjusted for key confounders to be considered for inclusion. Prognostic data were only pooled if studies were similar in terms of populations, risk factors, outcomes and statistical analysis methods (including adjustments for confounding factors), otherwise results from individual studies were presented in the evidence reports.

Data synthesis for qualitative reviews

Where possible, a meta-synthesis was conducted to combine evidence from qualitative studies. Whenever studies identified a qualitative theme relevant to the protocol, this was extracted and the main characteristics were summarised. When all themes had been extracted from studies, common concepts were categorised and tabulated. This included information on how many studies had contributed to each theme identified by the NGA technical team.

In qualitative synthesis, a theme being reported more than other themes across included studies does not necessarily mean that the theme is more important than other themes. The aim of qualitative research is to identify new perspectives on a particular topic. Study types and populations in qualitative research can differ widely, meaning that themes identified by just one or a few studies can provide important new information on a given topic.

Themes from individual studies were integrated into a wider context and, when possible, overarching categories of themes with sub-themes were identified. Themes were derived from data presented in individual studies. When themes were extracted from 1 primary study only, theme names used in the guideline mirrored those in the source study. However, when themes were based on evidence from multiple studies, the theme names were assigned by the NGA technical team. The names of overarching categories of themes were also assigned by the NGA technical team.

Emerging themes were placed into a thematic map representing the relationship between themes and overarching categories. The purpose of such a map is to show relationships between overarching categories and associated themes.

Appraising the quality of evidence

Intervention studies

GRADE methodology for intervention reviews

For intervention reviews, the evidence for outcomes from included RCTs and comparative non-randomised studies was evaluated and presented using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) methodology developed by the international GRADE working group.

When GRADE was applied, software developed by the GRADE working group (GRADEpro) was used to assess the quality of each outcome, taking account of individual study quality factors and any meta-analysis results. Results were presented in GRADE profiles (GRADE tables).

The selection of outcomes for each review question was agreed during development of the associated review protocol in discussion with the committee. The evidence for each outcome was examined separately for the quality elements summarised in Table 2. Criteria considered in the rating of these elements are discussed below. Each element was graded using the quality ratings summarised in Table 3. Footnotes to GRADE tables were used to record reasons for grading a particular quality element as having a ‘serious’ or ‘very serious’ quality issue. The ratings for each component were combined to obtain an overall assessment of quality for each outcome as described in Table 4.

The initial quality rating was based on the study design: RCTs start as 'high' quality evidence and non-randomised studies as 'low' quality evidence. The rating was then modified according to the assessment of each quality element (Table 2). Each quality element considered to have a 'serious' or 'very serious' quality issue was downgraded by 1 or 2 levels respectively (for example, evidence starting as 'high' quality was downgraded to 'moderate' or 'low' quality). In addition, there was a possibility to upgrade evidence from non-randomised studies (provided the evidence for that outcome had not previously been downgraded) if there was a large magnitude of effect, a dose–response gradient, or if all plausible confounding would reduce a demonstrated effect or suggest a spurious effect when results showed no effect.

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

Quality element	Description
Risk of bias ('Study limitations')	Limitations in study design and implementation may bias estimates of treatment effect. High risk of bias for the majority of the evidence reduces confidence in the estimated effect
Inconsistency	This refers to unexplained heterogeneity in the results
Indirectness	This refers to differences in study populations, interventions, comparators or outcomes between the available evidence and inclusion criteria specified in the review protocol
Imprecision	This occurs when a study has relatively few participants or few events of interest, resulting in wide confidence intervals around estimates of effect that include clinically important thresholds
Publication bias	This refers to systematic under- or over-estimation of the underlying benefit or harm resulting from selective publication of study results

Table 3: GRADE quality ratings (by quality element)

Quality issues	Description
None or not serious	No serious issues with the evidence for the quality element under consideration
Serious	Issues with the evidence sufficient to downgrade by 1 level for the quality element under consideration
Very serious	Issues with the evidence sufficient to downgrade by 2 levels for the quality element under consideration

Table 4: Overall quality of the evidence in GRADE (by outcome)

Overall quality grading	Description
High	Further research is very unlikely to change the level of confidence in the estimate of effect
Moderate	Further research is likely to have an important impact on the level of confidence in the estimate of effect and may change the estimate
Low	Further research is very likely to have an important impact on the level of confidence in the estimate of effect and is likely to change the estimate
Very low	The estimate of effect is very uncertain

Assessing risk of bias in intervention reviews

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

Risk of bias in RCTs was assessed using the Cochrane risk of bias tool (see Appendix H in Developing NICE guidelines: the manual; NICE 2014).

The Cochrane risk of bias tool assesses the following possible sources of bias:

- selection bias
- performance bias
- attrition bias
- detection bias
- reporting bias.

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 the chosen design and methodology will impact on the estimation of the intervention effect. For example for outcomes considered objective (survival, complete resection rate, local recurrence, sphincter preservation/permanent stoma and treatment-related mortality) we did not automatically downgrade for lack of blinding.

More details about the Cochrane risk of bias tool can be found in Section 8 of the Cochrane Handbook for Systematic Reviews of Interventions (Higgins 2011).

For systematic reviews the ROBIS checklist was used (see Appendix H in Developing NICE guidelines: the manual; NICE 2014).

For non-randomised studies the ROBINS-I checklist was used (see Appendix H in Developing NICE guidelines: the manual; NICE 2014).

When meta-analysis was performed the respective weights of the individual studies in the meta-analysis were considered when assessing the quality of the evidence. For example, when an individual study had a serious risk of bias but only accounted for, for example, 2% of the weight of the meta-analysis, the evidence was not downgraded for serious risk of bias.

Assessing inconsistency in intervention reviews

Inconsistency refers to unexplained heterogeneity in results of meta-analysis. When estimates of 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 truly applicable when statistical meta-analysis is conducted (that is, results from different studies are pooled). When outcomes were derived from a single study the rating ‘no serious inconsistency’ was used when assessing this domain, as per GRADE methodology (Santesso 2016).

Inconsistency was assessed visually by inspecting forest plots and observing whether there was considerable heterogeneity in the results of the meta-analysis. This was assessed by calculating the I-squared statistic for the meta-analysis with an I-squared value of more than 50% indicating considerable heterogeneity, and more than 80% indicating very serious heterogeneity. When considerable or very serious heterogeneity was observed, possible reasons were explored and subgroup analyses were performed as pre-specified in the review protocol where possible. In the case of

unexplained heterogeneity, sensitivity analyses were planned based on the quality of studies, eliminating studies at high risk of bias (in relation to randomisation, allocation concealment and blinding, and/or missing outcome data).

When considerable heterogeneity was present, the meta-analysis was re-run using the Der-Simonian and Laird method with a random effects model to provide a more conservative estimate of the effect.

When no plausible explanation for the heterogeneity could be found, the quality of the evidence was downgraded in GRADE for inconsistency.

Assessing indirectness in intervention reviews

Directness refers to the extent to which populations, interventions, comparisons and outcomes reported in the evidence are similar to those defined in the inclusion criteria for the review and was assessed by comparing the PICO elements in the studies to the PICO defined in the review protocol. Indirectness is important when such differences are expected to contribute to a difference in effect size, or may affect the balance of benefits and harms considered for an intervention.

Assessing imprecision and clinical importance in intervention reviews

Imprecision in GRADE methodology refers to uncertainty around the effect estimate and whether or not there is a clinically important difference between interventions (that is, whether the evidence clearly supports a particular recommendation or appears to be consistent with several candidate recommendations). Therefore, imprecision differs from other aspects of evidence quality because it is not concerned with whether the point estimate is accurate or correct (has internal or external validity). Instead, it is concerned with uncertainty about what the point estimate actually represents. 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 the procedure to be repeated. The larger the study, the smaller the 95% CI will be and the more certain the effect estimate.

Imprecision was assessed in the guideline evidence reviews by considering whether the width of the 95% CI of the effect estimate was relevant to decision making, considering each outcome independently. This is illustrated in Figure 1, which considers a positive outcome for the comparison of treatment ‘A’ versus treatment ‘B’. Three decision-making zones can be differentiated, bounded by the thresholds for clinical importance (minimally important differences; MIDs) for benefit and harm. The MID for harm for a positive outcome means the threshold at which treatment A is less effective than treatment B by an amount that is clinically important to people with the condition of interest (favours B).

When the CI of the effect estimate is wholly contained in 1 of the 3 zones there is no uncertainty about the size and direction of effect, therefore, the effect estimate is considered precise; that is, there is no imprecision.

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

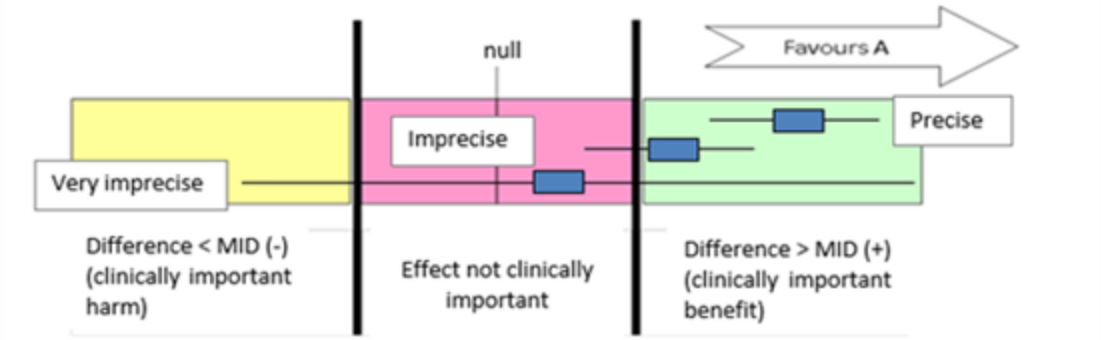
When the CI crosses all 3 zones, the effect estimate is considered to be very imprecise because the CI is consistent with 3 possible clinical decisions and there is therefore a considerable lack of confidence in the results. The evidence is therefore downgraded by 2 levels in the GRADE analysis ('very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

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

For most outcomes, however, there was a lack of an agreed MID. Instead outcomes with low numbers of events or small sample size were downgraded for imprecision using the following criteria

- For dichotomous outcomes: downgrade by 1 if <300 events altogether (across both arms)
- For continuous outcomes: downgrade by 1 if sample size <400 altogether (across both arms)

Figure 1: Assessment of imprecision and clinical importance in intervention reviews using GRADE



MID: minimally important difference

Defining minimally important differences for intervention reviews

The committee was asked whether there were any recognised or acceptable MIDs in the clinical literature and community relevant to the review questions under consideration. The following MIDs for quality of life measures were identified in the clinical literature and approved by the committee:

- EORTC QLQ-C30: 5 points
- EORTC QLQ-CR29: 5 points
- EORTC QLQ-CR38: 5 points
- EQ-5D: 0.09 using FACT-G quintiles
- FACT-C: 5 points
- FACT-G: 5 points*
- SF-12: > 3.77 for the mental component summary (MCS) and > 3.29 for the physical component summary (PCS) of the Short Form SF-12 (SF-12)
- SF-36: > 7.1 for the physical functioning scale, > 4.9 for the bodily pain scale, and > 7.2 for the physical component summary.

For other outcomes (for example, overall survival, disease-free survival or complication), any statistically significant change was considered by the committee to of potential clinically importance.

Prognostic studies

Adapted GRADE methodology for prognostic reviews

For prognostic reviews with evidence from comparative studies an adapted GRADE approach was used. As noted above, GRADE methodology is designed for intervention reviews but the quality assessment elements were adapted for prognostic reviews.

The evidence for each outcome in the prognostic reviews was examined separately for the quality elements listed and defined in Table 5. The criteria considered in the rating of these elements are discussed below. Each element was graded using the quality levels summarised in Table 3. Footnotes to GRADE tables were used to record reasons for grading a particular quality element as having 'serious' or 'very serious' quality issues. The ratings for each component were combined to obtain an overall assessment of quality for each outcome as described in Table 4.

Table 5: Adaptation of GRADE quality elements for prognostic reviews

Quality element	Description
Risk of bias ('Study limitations')	Limitations in study design and implementation may bias estimates and interpretation of the effect of the prognostic/risk factor. High risk of bias for the majority of the evidence reduces confidence in the estimated effect. Prognostic studies are not usually randomised and therefore would not be downgraded for study design from the outset (they start as high quality)
Inconsistency	This refers to unexplained heterogeneity between studies looking at the same prognostic/risk factor, resulting in wide variability in estimates of association (such as RRs or ORs), with little or no overlap in confidence intervals
Indirectness	This refers to any departure from inclusion criteria listed in the review protocol (such as differences in study populations or prognostic/risk factors), that may affect the generalisability of results
Imprecision	This occurs when a study has relatively few participants, a low event rate and also when the number of participants is too small for a multivariable analysis (as a rule of thumb, 10 participants are needed per variable).

RR: relative risk; OR: odds ratio

Assessing risk of bias in prognostic reviews

The Quality in Prognosis Studies (QUIPS) tool developed by Hayden 2013 was used to assess risk of bias in studies included in prognostic reviews (see Appendix H in the Developing NICE guidelines: the manual; NICE 2014). The risk of bias in each study was determined by assessing the following domains:

- selection bias
- attrition bias
- prognostic factor bias

- outcome measurement bias
- control for confounders
- appropriate statistical analysis.

Assessing inconsistency in prognostic reviews

Where multiple results were deemed appropriate to meta-analyse (i.e. sufficient similarity between risk factor and outcome under investigation) inconsistency was assessed by visually inspecting forest plots and observing whether there was considerable heterogeneity in the results of the meta-analysis. This was assessed by calculating the I-squared statistic for the meta-analysis with an I-squared value of more than 50% indicating considerable heterogeneity, and more than 80% indicating very serious heterogeneity. When considerable or very serious heterogeneity was observed, possible reasons were explored and subgroup analyses were performed as pre-specified in the review protocol where possible.

When no plausible explanation for the heterogeneity could be found, the quality of the evidence was downgraded in GRADE for inconsistency.

If meta-analysis was not appropriate (for example due to clinical heterogeneity) inconsistency was assessed by visually inspecting forest plots.

Assessing indirectness in prognostic reviews

Indirectness in prognostic reviews was assessed by comparing the populations, prognostic factors and outcomes in the evidence to those defined in the review protocol.

Assessing imprecision and clinical importance in prognostic reviews

The committee considered any statistically significant change to be of potential clinical importance. Outcomes were downgraded for imprecision if the overall sample size was considered too small for multivariable analysis (using a rule of thumb of 10 participants needed per included variable) or if there were fewer than 300 events for dichotomous outcomes.

Qualitative reviews

Adapted GRADE-CERQual methodology for qualitative reviews

For the qualitative review an adapted GRADE Confidence in the Evidence from Reviews of Qualitative research (GRADE-CERQual) approach (Lewin 2015) was used. In this approach the quality of evidence is considered according to themes in the evidence. The themes may have been identified in the primary studies or they may have been identified by considering the reports of a number of studies. Quality elements assessed using GRADE-CERQual are listed and defined in Table 6. Each element was graded using the levels of concern summarised in Table 7. The ratings for each component were combined (as with other types of evidence) to obtain an overall assessment of quality for each theme as described in Table 8.

Table 6: Adaptation of GRADE quality elements for qualitative reviews

Quality element	Description
Risk of bias ('Methodological limitations')	Limitations in study design and implementation may bias interpretation of qualitative themes identified. High risk of bias for the majority of the evidence reduces confidence in review findings. Qualitative studies are not usually randomised and therefore would not be downgraded for study design from the outset (they start as high quality)
Relevance (or applicability) of evidence	This refers to the extent to which the evidence supporting the review findings is applicable to the context specified in the review question
Coherence of findings	This refers to the extent to which review findings are well grounded in data from the contributing primary studies and provide a credible explanation for patterns identified in the evidence
Adequacy of data (theme saturation or sufficiency)	This corresponds to a similar concept in primary qualitative research, that is, 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 the particular theme. Individual studies that may have contributed to a theme or sub-theme may have been conducted in a manner that by design would have not reached theoretical saturation at an individual study level

Table 7: CERQual levels of concern (by quality element)

Level of concern	Definition
None or very minor concerns	Unlikely to reduce confidence in the review finding
Minor concerns	May reduce confidence in the review finding
Moderate concerns	Will probably reduce confidence in the review finding
Serious concerns	Very likely to reduce confidence in the review finding

Table 8: Overall confidence in the evidence in CERQual (by review finding)

Overall confidence level	Definition
High	It is highly likely that the review finding is a reasonable representation of the phenomenon of interest
Moderate	It is likely that the review finding is a reasonable representation of the phenomenon of interest
Low	It is possible that the review finding is a reasonable representation of the phenomenon of interest
Very low	It is unclear whether the review finding is a reasonable representation of the phenomenon of interest

Assessing methodological limitations in qualitative reviews

Methodological limitations in qualitative studies were assessed using the Critical Appraisal Skills Programme (CASP) checklist for qualitative studies (see appendix H in Developing NICE guidelines: the manual; NICE 2014). Overall methodological limitations was derived by assessing the 6 domains summarised in Table 9.

Table 9: Methodological limitations in qualitative studies

Aim and appropriateness of qualitative evidence	This domain assesses whether the aims and relevance of the study were described clearly 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 was documented clearly and whether it was based on a theoretical framework (such as ethnography or grounded theory). This does not necessarily mean that the framework has to be stated explicitly, but a detailed description ensuring transparency and reproducibility should be provided
Sample selection	This domain assesses the background, the procedure and reasons for the method of selecting participants. The assessment should include consideration of any relationship between the researcher and the participants, and how this might have influenced the findings
Data collection	This domain assesses the documentation of the method of data collection (in-depth interviews, semi-structured interviews, focus groups or observations). It also assesses who conducted any interviews, how long they lasted and where they took place
Data analysis	This domain assesses whether sufficient detail was documented for the analytical process and whether it was in accordance with the theoretical approach. For example, if a thematic analysis was used, the assessment would focus on the description of the approach used to generate themes. Consideration of data saturation would also form part of this assessment (it could be reported directly or it might be inferred from the citations documented that more themes could be found)
Results	This domain assesses any reasoning accompanying reporting of results (for example, whether a theoretical proposal or framework is provided)

Assessing relevance of evidence in qualitative reviews

Relevance (applicability) of findings in qualitative research is the equivalent of indirectness for quantitative outcomes, and refers to how closely the aims and context of studies contributing to a theme reflect the objectives outlined in the guideline review protocol.

Assessing coherence of findings in qualitative reviews

For qualitative research, a similar concept to inconsistency is coherence, which refers to the way findings within themes are described and whether they make sense. This concept was used in the quality assessment across studies for individual themes. This does not mean that contradictory evidence was automatically downgraded, but that it was highlighted and presented, and that reasoning was provided. Provided the themes, or components of themes, from individual studies fit into a theoretical framework, they do not necessarily have to reflect the same perspective. It should, however, be possible to explain these by differences in context (for example, the views of healthcare professionals might not be the same as those of family members, but they could contribute to the same overarching themes).

Assessing adequacy of data in qualitative reviews

Adequacy of data (theme saturation or sufficiency) corresponds to a similar concept in primary qualitative research in which consideration is made of whether a theoretical point of theme saturation was achieved, meaning that no further citations or observations would provide more insight or suggest a different interpretation of the theme concerned. As noted above, it is not equivalent to the number of studies contributing to a theme, but rather to the depth of evidence and whether sufficient quotations or observations were provided to underpin the findings.

Assessing clinical importance in qualitative reviews

For themes stemming from qualitative findings, clinical importance was agreed by the committee taking account of the generalisability of the context from which the theme was derived and whether it was sufficiently convincing to support or warrant a change in current practice, as well as the quality of the evidence.

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 10.

Table 10: 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 quality of evidence was assessed using the economic evaluations checklist as specified in Developing NICE guidelines: the manual (NICE 2014).

Appraising the quality of economic evidence

The quality of economic evaluations in this guideline were appraised using the methodology checklist reported in the Developing NICE guidelines: the manual, Appendix H (NICE 2014) 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 colorectal cancer and colorectal cancer 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 2 economic models on: the treatment of liver metastases which were amenable to treatment with curative intent and surgical techniques for rectal cancer. The committee thought economic considerations would be particularly important when formulating recommendations in these areas.

The methods and results of the de novo economic analyses are reported in Appendix J of the relevant evidence reports. 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 important 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.

Expert evidence

When the guideline developers believe that there is important evidence in addition to that identified by the literature searches, they can call upon relevant individuals or organisations to provide expert evidence to the guideline committee.

When agreeing the review protocols, the committee considered areas that might benefit from expert evidence. The committee were aware of 2 randomised trials due to publish just before the guideline itself but not soon enough to be included in the evidence review. They agreed that the results of these trials could change current practice and invited the trialists to provide expert evidence. In both cases this evidence was provided orally and was discussed and considered by the committee. A written summary of the expert evidence was also included as an appendix to both evidence reports.

The topics covered by expert evidence were:

- neoadjuvant systemic anticancer therapy for operable colon cancer
- stenting for large bowel obstruction caused by colorectal cancer.

A call for evidence to all registered stakeholders was not made because the committee did not believe there was relevant unpublished evidence other areas, beyond the two topics identified above. Unpublished evidence in the form of conference abstracts was not included because such abstracts typically do not have enough detail to be critical appraised.

Developing recommendations

Guideline recommendations

Recommendations were drafted on the basis of the committee's interpretation of the available evidence, taking account of 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 their expert opinion. The considerations for making consensus-based recommendations include the balance between potential benefits and harms, the economic costs or implications compared with the economic benefits, current practices, recommendations made in other relevant guidelines, preferences of people and equality issues.

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

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

Research recommendations

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

Validation process

This guideline was subject to a 6-week public consultation and feedback process. All comments received from registered stakeholders were responded to in writing and posted on the NICE website at publication. For further details refer to Developing NICE guidelines: the manual (NICE 2014).

Updating the guideline

Following publication, NICE will undertake a surveillance review to determine whether the evidence base has progressed sufficiently to consider altering the guideline recommendations and warrant an update. For further details refer to Developing NICE guidelines: the manual (NICE 2014).

Funding

The NGA was commissioned by NICE to develop this guideline.

References

Bradburn 2007

Bradburn M, Deeks J, Berlin J, et al. (2007) Much ado about nothing: A comparison of the performance of meta-analytical methods with rare events. Statistics in Medicine 26(1): 53–77

Hayden 2013

Hayden J, van der Windt D, Cartwright L, et al. (2013) Assessing Bias in Studies of Prognostic Factors. Annals of Internal Medicine 158(4): 280–286

Higgins 2011

Higgins J, Green S (editors) (2011) Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011] The Cochrane Collaboration. Available from www.handbook.cochrane.org (accessed 15 June 2018)

Lewin 2015

Lewin S, Glenton C, Munthe-Kaas, H et al. (2015) Using qualitative evidence in decision making for health and social interventions: an approach to assess confidence in findings from qualitative evidence syntheses (GRADE-CERQual). PLoS Med 12(10): e1001895

NICE 2014

National Institute for Health and Care Excellence (NICE) (2019) Developing NICE guidelines: the manual (updated 2018, accessed 15 June 2018)

NICE 2018

National Institute for Health and Care Excellence (NICE) (2014) NICE Policy on conflict of interest (updated 2017, accessed 15 June 2018)

Santesso 2016

Santesso N, Carrasco-Labra A, Langendam M, et al. (2016) 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

Tierney 2007

Tierney J, Stewart L, Ghersi D, et al. (2007) Practical methods for incorporating summary time-to-event data into meta-analysis. Trials 8: 1-16

Viechtbauer 2010

Viechtbauer W (2010). Conducting meta-analyses in R with the metafor package. Journal of Statistical Software 36(3): 1-48