National Institute for Health and Care Excellence

Final

Diabetic retinopathy: management and monitoring

NICE guideline: methods

NICE guideline NG242

Methods

August 2024

Final

Evidence reviews were developed by NICE

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 All NICE guidance is subject to regular review and may be updated or withdrawn.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: 978-1-4731-6440-6

Contents

Development of the guideline

Remit

To see “What this guideline covers” and “What this guideline does not cover” please see the diabetic retinopathy guideline scope.

Methods

This guideline was developed using the methods described in the 2022 NICE guidelines manual.

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

Developing the review questions and outcomes

The 12 review questions developed for this guideline were based on the key areas identified in the guideline scope. They were drafted by the NICE guideline development team and refined and validated by the guideline committee.

The review questions were based on the following frameworks:

- Population, Intervention, Comparator and Outcome [and Study type] (PICO[S]) for reviews of interventions
- Population, index test(s), reference standard and outcome for reviews of diagnostic and predictive accuracy

Evidence reviews were completed for all review questions.

Reviewing research evidence

Review protocols

Review protocols were developed with the guideline committee to outline the inclusion and exclusion criteria used to select studies for each evidence review. Where possible, review protocols were prospectively registered in the PROSPERO register of systematic reviews.

Searching for evidence

Evidence was searched for each review question using the methods specified in the 2022 NICE guidelines manual.

Selecting studies for inclusion

All references identified by the literature searches and from other sources (for example, previous versions of the guideline or studies identified by committee members) were uploaded into EPPI reviewer software (version 5) and de-duplicated. Titles and abstracts were assessed for possible inclusion using the criteria specified in the review protocol. 10% of the abstracts were reviewed by two reviewers, with any disagreements resolved by discussion or, if necessary, a third independent reviewer.

The following evidence reviews made use of the priority screening functionality within the EPPI-reviewer software: Evidence review B: Effectiveness of different thresholds or criteria for starting treatment for non-proliferative diabetic retinopathy, proliferative

diabetic retinopathy, and diabetic macular oedema and Evidence review C: effectiveness of intensive treatments to lower blood glucose levels on progression of diabetic retinopathy and diabetic macular oedema. This functionality uses a machine learning algorithm (specifically, an SGD classifier) to take information on features (1, 2 and 3 word blocks) in the titles and abstract of papers marked as being 'includes' or 'excludes' during the title and abstract screening process, and re-orders the remaining records from most likely to least likely to be an include, based on that algorithm. This re-ordering of the remaining records occurs every time 25 additional records have been screened. Research is currently ongoing as to what are the appropriate thresholds where reviewing of abstracts can be stopped, assuming a defined threshold for the proportion of relevant papers it is acceptable to miss on primary screening. As a conservative approach until that research has been completed, the following rules were adopted during the production of this guideline:

- In every review, at least 50% of the identified abstracts were always screened.
- After this point, screening was only terminated if an additional 5% of the database was screened without finding an include based on title and abstract screening.

These stopping criteria were considered appropriate based on the experience of the team, given this topic is a well-defined clinical area with clear inclusion and exclusion criteria. As additional measure, it was specified that the full database would be searched if there were a very small number of included studies (<30).

As an additional check to ensure this approach did not miss relevant studies, committee members were consulted to identify studies that were missed. If systematic reviews (or qualitative evidence syntheses in the case of reviews of qualitative studies) were included in the review protocol, relevant systematic reviews or qualitative evidence syntheses were used to identify any papers not found through the primary search. If additional studies were found that were erroneously excluded during the priority screening process, the full database was subsequently screened.

The decision whether or not to use priority screening was taken by the reviewing team depending on the perceived likelihood that stopping criteria would be met, based on the size of the database, heterogeneity of studies included in the review and predicted number of includes. If it was thought that stopping criteria were unlikely to be met, priority screening was not used, and the full database was screened.

The full text of potentially eligible studies was retrieved and assessed according to the criteria specified in the review protocol. A standardised form was used to extract data from included studies.

Incorporating published evidence syntheses

If published evidence syntheses were identified sufficiently early in the review process (for example, from the surveillance review or early in the database search), they were considered for use as the primary source of data, rather than extracting information from primary studies. Syntheses considered for inclusion in this way were quality assessed to assess their suitability using the appropriate checklist, as outlined in Table 1. Note that this quality assessment was solely used to assess the quality of the synthesis in order to decide whether it could be used as a source of data, as outlined in Table 2, not the quality of evidence contained within it, which was

assessed in the usual way as outlined in the section on 'Appraising the quality of evidence'.

Table 1: Checklists for published evidence syntheses

Type of synthesis	Checklist for quality appraisal
Systematic review of quantitative evidence	ROBIS
Network meta-analysis	Modified version of the PRISMA NMA tool (see appendix K of 'Developing NICE guidelines, the manual')
Qualitative evidence synthesis	ENTREQ reporting standard for published evidence synthesis (https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-12-181) is the generic reporting standard for QES, however specific reporting standards exist for meta-ethnography (eMERGe [https://emergeproject.org/]) and for realist synthesis (RAMESES II [https://www.ramesesproject.org/]). If these reporting standards are not appropriate to the QES then an adapted PRISMA framework is used (see Flemming K, Booth A, Hannes K, Cargo M, Noyes J. Cochrane Qualitative and Implementation Methods Group guidance series-paper 6: reporting guidelines for qualitative, implementation, and process evaluation evidence syntheses. Journal of Clinical Epidemiology 2018; 97: 79-85).
Individual patient data meta-analysis	Checklist based on Tierney, Jayne F., et al. "Individual participant data (IPD) meta-analyses of randomised controlled trials: guidance on their use." PLoS Med 12.7 (2015): e1001855.

Each published evidence synthesis was classified into one of the following three groups:

- High quality – It is unlikely that additional relevant and important data would be identified from primary studies compared to that reported in the review, and unlikely that any relevant and important studies have been missed by the review.
- Moderate quality – It is possible that additional relevant and important data would be identified from primary studies compared to that reported in the review, but unlikely that any relevant and important studies have been missed by the review.
- Low quality – It is possible that relevant and important studies have been missed by the review.

Each published evidence synthesis was also classified into one of three groups for its applicability as a source of data, based on how closely the review matches the specified review protocol in the guideline. Studies were rated as follows:

- Fully applicable – The identified review fully covers the review protocol in the guideline.
- Partially applicable – The identified review fully covers a discrete subsection of the review protocol in the guideline (for example, some of the factors in the protocol only).
- Not applicable – The identified review, despite including studies relevant to the review question, does not fully cover any discrete subsection of the review protocol in the guideline.

The way that a published evidence synthesis was used in the evidence review depended on its quality and applicability, as defined in Table 2. When published evidence syntheses were used as a source of primary data, data from these evidence syntheses were quality assessed and presented in GRADE/CERQual tables in the same way as if data had been extracted from primary studies. In questions where data was extracted from both systematic reviews and primary studies, these were checked to ensure none of the data had been double counted through this process.

Table 2: Criteria for using published evidence syntheses as a source of data

Quality	Applicability	Use of published evidence synthesis
High	Fully applicable	Data from the published evidence synthesis were used instead of undertaking a new literature search or data analysis. Searches were only done to cover the period of time since the search date of the review. If the review was considered up to date (following discussion with the guideline committee and NICE lead for quality assurance), no additional search was conducted.
High	Partially applicable	Data from the published evidence synthesis were used instead of undertaking a new literature search and data analysis for the relevant subsection of the protocol. For this section, searches were only done to cover the period of time since the search date of the review. If the review was considered up to date (following discussion with the guideline committee and NICE lead for quality assurance), no additional search was conducted. For other sections not covered by the evidence synthesis, searches were undertaken as normal.
Moderate	Fully applicable	Details of included studies were used instead of undertaking a new literature search. Full-text papers of included studies were still retrieved for the purposes of data analysis. Searches were only done to cover the period of time since the search date of the review.
Moderate	Partially applicable	Details of included studies were used instead of undertaking a new literature search for the relevant subsection of the protocol. For this section, searches were only done to cover the period of time since the search date of the review. For other sections not covered by the evidence synthesis, searches were undertaken as normal.

Methods of combining evidence

Data synthesis for intervention studies

Where possible, meta-analyses were conducted to combine the results of quantitative studies for each outcome. Network meta-analyses was considered in situations where there were at least 3 treatment alternatives. When there were 2 treatment alternatives, pairwise meta-analysis was used to compare interventions.

Pairwise meta-analysis

Pairwise meta-analyses were performed in Cochrane Review Manager V5.3. A pooled relative risk was calculated for dichotomous outcomes (using the Mantel–Haenszel method) reporting numbers of people having an event. Both relative and absolute risks were presented, with absolute risks calculated by applying the relative risk to the risk in the comparator arm of the meta-analysis (calculated as the total number events in the comparator arms of studies in the meta-analysis divided by the total number of participants in the comparator arms of studies in the meta-analysis).

A pooled mean difference was calculated for continuous outcomes (using the inverse variance method) when the same scale was used to measure an outcome across different studies. Where different studies presented continuous data measuring the same outcome but using different numerical scales (e.g. a 0-10 and a 0-100 visual analogue scale), these outcomes were all converted to the same scale before meta-analysis was conducted on the mean differences.

For continuous outcomes analysed as mean differences, change from baseline values were used in the meta-analysis if they were accompanied by a measure of spread (for example standard deviation). Where change from baseline (accompanied by a measure of spread) were not reported, the corresponding values at the timepoint of interest were used.

Random effects models were fitted when significant between-study heterogeneity in methodology, population, intervention or comparator was identified by the reviewer in advance of data analysis. This decision was made and recorded before any data analysis was undertaken. For all other syntheses, fixed- and random-effects models were fitted, with the presented analysis dependent on the degree of heterogeneity in the assembled evidence. Fixed-effects models were the preferred choice to report, but in situations where the assumption of a shared mean for fixed-effects model were clearly not met, even after appropriate pre-specified subgroup analyses were conducted, random-effects results are presented. Fixed-effects models were deemed to be inappropriate if there was significant statistical heterogeneity in the meta-analysis, defined as $I^2 \geq 50\%$.

However, in cases where the results from individual pre-specified subgroup analyses were less heterogeneous (with $I^2 < 50\%$) the results from these subgroups were reported using fixed effects models. This may have led to situations where pooled results were reported from random-effects models and subgroup results were reported from fixed-effects models.

Network meta-analysis

We undertook hierarchical Bayesian network meta-analysis using WinBUGS version 1.4.3. The models used reflected the recommendations of the NICE Decision Support Unit's Technical Support Documents (TSDs) on evidence synthesis, particularly TSD 2 ('A generalised linear modelling framework for pairwise and network meta-analysis of randomised controlled trials'; see http://www.nicedsu.org.uk/). We used the WinBUGS code provided in the appendices of TSD 2 without substantive alteration to specify synthesis models. We used a normal likelihood with correction for multi-arm trials. Non-informative prior distributions were used for all parameters. Priors were normally distributed with a mean of 0 and variance of 10,000, except for the standard deviation between trials for the random effects meta-analyses which had a uniform prior distribution ranging

from 0 to 5 for the visual acuity outcomes and from 0 to 1000 for the central retinal thickness outcomes. Standard threshold laser treatment was used as the reference treatment as this treatment has a high number of links with other nodes in the network and is commonly used as first-line treatment.

Results were assessed for convergence to determine the length of 'burn in' period required by examining the 'bgdiag' and 'history' plots. The MC error was assessed to check that it was sufficiently small (less than 5% of the standard deviation of the posterior distribution for each parameter) and additional samples were summarised if this was not the case. We report results summarising 50,000 samples from the posterior distribution of each model, having first run and discarded 50,000 'burn-in' iterations. Three separate chains with different initial values were used.

Fixed - and random-effects models were explored for each outcome, with the final choice of model based on the total residual deviance and deviance information criterion (DIC). The total residual deviance reflects the model's ability to predict the individual data points underlying it – a well-fitting model will have a total residual deviance approximately equal to the number of data points. DIC provides an estimate of deviance that is 'penalised' according to the number of parameters in the model, and is useful for comparing models on the same dataset. If DIC was at least 3 points lower for the random-effects model, it was preferred; otherwise, the fixed effects model was considered to provide an equivalent fit to the data in a more parsimonious analysis, and was preferred.

Inconsistency between direct and indirect evidence was assessed when possible by fitting 'inconsistency models' to the data and assessing model fit using the deviance information criteria. A reduction in DIC of 3 or more was taken as evidence of possible inconsistency. For random effect models, the between studies standard deviation was also inspected, with a fall in the between-studies standard deviation taken as evidence of inconsistency. If inconsistency was identified, the source of this inconsistency was explored and resolved if possible (for example by re-evaluating which studies are included in the network). If inconsistency could not be resolved then this was reflected in the quality assessment for the network meta-analysis (see Modified GRADE for intervention studies analysed using network meta-analysis),

Data synthesis for diagnostic accuracy data

In this guideline, diagnostic test accuracy (DTA) data are classified as any data in which a feature – be it a symptom, a risk factor, a test result or the output of some algorithm that combines many such features – is observed in some people who have the condition of interest at the time of the test and some people who do not. Such data either explicitly provide, or can be manipulated to generate, a 2x2 classification of true positives and false negatives (in people who, according to the reference standard, truly have the condition) and false positives and true negatives (in people who, according to the reference standard, do not).

The 'raw' 2x2 data can be summarised in a variety of ways. Those that were used for decision making in this guideline were as follows:

- **Positive likelihood ratios** describe how many times more likely positive features are in people with the condition compared to people without the condition. Values greater than 1 indicate that a positive result makes the condition more likely.
  - LR^+ = (TP/[TP+FN])/(FP/[FP+TN])

- **Negative likelihood ratios** describe how many times less likely negative features are in people with the condition compared to people without the condition. Values less than 1 indicate that a negative result makes the condition less likely.
  - LR^- = (FN/[TP+FN])/(TN/[FP+TN])
- **Sensitivity** is the probability that the feature will be positive in a person with the condition.
  - sensitivity = TP/(TP+FN)
- **Specificity** is the probability that the feature will be negative in a person without the condition.
  - specificity = TN/(FP+TN)

Meta-analysis of diagnostic accuracy data was conducted with reference to the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 2.1 (Deeks et al. 2022).

Where five or more studies were available for all included strata, a bivariate model was fitted using the `mada` package in R v3.4.0, which accounts for the correlations between positive and negative likelihood ratios, and between sensitivities and specificities. Where sufficient data were not available (2-4 studies), separate independent pooling was performed for positive likelihood ratios, negative likelihood ratios, sensitivity and specificity, using R. This approach is conservative as it is likely to somewhat underestimate test accuracy, due to failing to account for the correlation and trade-off between sensitivity and specificity (see Deeks 2010).

Random-effects models (der Simonian and Laird) were fitted for all syntheses, as recommended in the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy (Deeks et al. 2010).

Data synthesis for association data

In this guideline, association data were defined as measures of association between one or more factors (which could be either a single variable or a group of variables) and an outcome variable, where the data are not reported in terms of outcome classification (i.e. diagnostic/predictive accuracy). Examples could include (but were not limited to) data assessing the association between variables and diagnosis (diagnostic association studies) or data assessing the association between variables and a future outcome (prognostic association studies). Data were reported as hazard ratios (if measured over time) or odds ratios or risk ratios (if measured at a specific time-point).

The same methods for meta-analysis of odds ratios and relative risks were used as described as in the section on Data synthesis for intervention studies.

Where appropriate, hazard ratios were pooled using the generic inverse-variance method. Adjusted odds ratios, hazard ratios and risk ratios from multivariate models were only pooled if the same set of factors were used across multiple studies and if the same thresholds to measure factors were used across studies.

Appraising the quality of evidence

Intervention studies (relative effect estimates)

RCTs and quasi-randomised controlled trials were quality assessed using the Cochrane Risk of Bias Tool. Non-randomised controlled trials and cohort studies were quality assessed using the ROBINS-I tool. Other study types (for example controlled before and after studies) were assessed using the preferred option specified in the NICE guidelines manual 2018 (appendix H). Evidence on each outcome for each individual study was classified into one of the following groups:

- Low risk of bias – The true effect size for the study is likely to be close to the estimated effect size.
- Moderate risk of bias – There is a possibility the true effect size for the study is substantially different to the estimated effect size.
- High risk of bias – It is likely the true effect size for the study is substantially different to the estimated effect size.
- Critical risk of bias (ROBINS-I only) - It is very likely the true effect size for the study is substantially different to the estimated effect size.

Each individual study was also classified into one of three groups for directness, based on if there were concerns about the population, intervention, comparator and/or outcomes in the study and how directly these variables could address the specified review question. Studies were rated as follows:

- Direct – No important deviations from the protocol in population, intervention, comparator and/or outcomes.
- Partially indirect – Important deviations from the protocol in one of the following areas: population, intervention, comparator and/or outcomes.
- Indirect – Important deviations from the protocol in at least two of the following areas: population, intervention, comparator and/or outcomes.

Minimally important differences (MIDs) and clinical decision thresholds

The Core Outcome Measures in Effectiveness Trials (COMET) database was searched to identify published minimal clinically important difference thresholds relevant to this guideline that might aid the committee in identifying clinical decision thresholds. Identified MIDs were assessed to ensure they had been developed and validated in a methodologically rigorous way, and were applicable to the populations, interventions and outcomes specified in this guideline. In addition, the Guideline Committee were asked to prospectively specify any outcomes where they felt a consensus clinical decision threshold could be defined from their experience. In particular, any questions looking to evaluate non-inferiority (that one treatment is not meaningfully worse than another) required a clinical decision threshold to be defined to act as a non-inferiority margin.

Clinical decision thresholds were used to aid interpretation of the size of effects for different outcomes. Clinical decision threshold that were used in the guideline are given in Table 3 and also reported in the relevant evidence reviews.

Table 3: Identified Clinical decision thresholds

Outcome	Clinical decision threshold	Source
Visual acuity	10 letters on ETDRS chart	Rosser DA, Cousens SN, Murdoch IE, Fitzke FW, Laidlaw DA. How sensitive to clinical change are ETDRS logMAR visual acuity measurements? Invest Ophthalmol Vis Sci. 2003;44(8):3278–3281.
Visual function question 25	6 points (for a population with glaucoma, thought to be applicable to a population with proliferative diabetic retinopathy as they would have similar baseline vision loss)	Burr JM, Cooper D, Ramsay CR, Che Hamzah J, Azuara-Blanco A. Interpretation of change scores for the National Eye Institute Visual Function Questionnaire-25: the minimally important difference. Br J Ophthalmol. 2021 May 18:bjophthalmol-2021-318901. doi: 10.1136/bjophthalmol-2021-318901. Epub ahead of print. PMID: 34006510.
Visual function questionnaire-25	3.33 points (for a population with diabetic macular oedema)	Lloyd AJ, Loftus J, Turner M, Lai G, Pleil A. Psychometric validation of the Visual Function Questionnaire-25 in patients with diabetic macular edema. Health Qual Life Outcomes. 2013;11:10.

This evidence reviews for this guideline was conducted using a modified version of the GRADE approach to rate the certainty of evidence in systematic reviews. Instead of using predefined MIDs to assess imprecision in GRADE profiles, imprecision was assessed qualitatively during committee discussions. These discussions involved consideration of published MIDs where they exist, but the committee were also encouraged to make judgements of imprecision based on the 95% confidence intervals and sample sizes reported in the GRADE tables. The committee were not aware of any published MIDs for any of the outcomes in the intervention reviews and so the discussions were based on the width of confidence intervals and whether they crossed the line of no effect. This should enable judgements of clinical importance to be made in the context of wider decision making, taking into account evidence across all outcomes and analyses, including health economic analyses.

Committee discussions regarding the clinical importance of effects was recorded in the ‘imprecision and clinical importance of effects’ section of the evidence review. In particular, this included consideration of whether the whole effect of a treatment (which may be felt across multiple independent outcome domains) would be likely to be clinically meaningful, rather than simply whether each individual sub outcome might be meaningful in isolation. The impact of imprecision on the recommendations was presented in the ‘quality of the evidence’ section of the committee discussion in the evidence review.

GRADE for intervention studies analysed using pairwise analysis

GRADE was used to assess the quality of evidence for the outcomes specified in the review protocol. Data from randomised controlled trials, non-randomised controlled trials and cohort studies (which were quality assessed using the Cochrane risk of bias tool or ROBINS-I) were initially rated as high quality while data from other study types were initially rated as low quality. The quality of the evidence for each outcome was downgraded or not from this initial point, based on the criteria given in Table 4. These criteria were used to apply preliminary ratings, but were overridden in cases

where, in the view of the analyst or committee the uncertainty identified was unlikely to have a meaningful impact on decision making.

Table 4: Rationale for downgrading quality of evidence for intervention studies

GRADE criteria	Reasons for downgrading quality
Risk of bias	Not serious: If less than 33.3% of the weight in a meta-analysis came from studies at moderate or high risk of bias, the overall outcome was not downgraded. Serious: If greater than 33.3% of the weight in a meta-analysis came from studies at moderate or high risk of bias, the outcome was downgraded one level. Very serious: If greater than 33.3% of the weight in a meta-analysis came from studies at high risk of bias, the outcome was downgraded two levels. Extremely serious: If greater than 33.3% of the weight in a meta-analysis came from studies at critical risk of bias, the outcome was downgraded three levels
Indirectness	Not serious: If less than 33.3% of the weight in a meta-analysis came from partially indirect or indirect studies, the overall outcome was not downgraded. Serious: If greater than 33.3% of the weight in a meta-analysis came from partially indirect or indirect studies, the outcome was downgraded one level. Very serious: If greater than 33.3% of the weight in a meta-analysis came from indirect studies, the outcome was downgraded two levels.
Inconsistency	Concerns about inconsistency of effects across studies, occurring when there is unexplained variability in the treatment effect demonstrated across studies (heterogeneity), after appropriate pre-specified subgroup analyses have been conducted. This was assessed using the I^2 statistic. N/A: Inconsistency was marked as not applicable if data on the outcome was only available from one study. Not serious: If the I^2 was less than 33.3%, the outcome was not downgraded. Serious: If the I^2 was between 33.3% and 66.7%, the outcome was downgraded one level. Very serious: If the I^2 was greater than 66.7%, the outcome was downgraded two levels.
Imprecision	This was not included in the GRADE table, but was considered during committee discussions of the evidence, taking into account 95% confidence intervals around the point estimate of the effect, any relevant MIDs, committee expertise and the effect of a single intervention based on multiple outcomes.
Publication bias	Where 10 or more studies were included as part of a single meta-analysis, a funnel plot was produced to graphically assess the potential for publication bias. When a funnel plot showed convincing evidence of publication bias, or the review team became aware of other evidence of publication bias (for example, evidence of unpublished trials where there was evidence that the effect estimate differed in published and unpublished data), the outcome was downgraded once. If no evidence of publication bias was found for any outcomes in a review (as was often the case), this domain was excluded from GRADE profiles to improve readability.

GRADE criteria	Reasons for downgrading quality

For outcomes that were originally assigned a quality rating of 'low' (when the data was from observational studies that were not appraised using the ROBINS-I checklist), the quality of evidence for each outcome was upgraded if any of the following three conditions were met and the risk of bias for the outcome was rated as 'no serious':

- Data from studies showed an effect size sufficiently large that it could not be explained by confounding alone.
- Data showed a dose-response gradient.
- Data where all plausible residual confounding was likely to increase our confidence in the effect estimate.

Modified GRADE for intervention studies analysed using network meta-analysis

A modified version of the standard GRADE approach for pairwise interventions was used to assess the quality of evidence across the network meta-analyses. While most criteria for pairwise meta-analyses still apply, it is important to adapt some of the criteria to take into consideration additional factors, such as how each 'link' or pairwise comparison within the network applies to the others. As a result, the following was used when modifying the GRADE framework to a network meta-analysis. It is designed to provide a single overall quality rating for an NMA to judge the overall strength of evidence. Additionally, where appropriate, threshold analysis was considered to explore the uncertainties within the NMA at contrast level.

These criteria were used to apply preliminary ratings, but were overridden in cases where, in the view of the analyst or committee the uncertainty identified was unlikely to have a meaningful impact on decision making.

Table 5: Rationale for downgrading quality of evidence for network meta-analysis

GRADE criteria	Reasons for downgrading quality
Risk of bias	Not serious: If fewer than 33.3% of the studies in the network meta-analysis were at moderate or high risk of bias, the overall network was not downgraded. Serious: If greater than 33.3% of the studies in the network meta-analysis were at moderate or high risk of bias, the network was downgraded one level. Very serious: If greater than 33.3% of the studies in the network meta-analysis were at high risk of bias, the network was downgraded two levels.
Indirectness	Not serious: If fewer than 33.3% of the studies in the network meta-analysis were partially indirect or indirect, the overall network was not downgraded. Serious: If greater than 33.3% of the studies in the network meta-analysis were partially indirect or indirect, the network was downgraded one level. Very serious: If greater than 33.3% of the studies in the network meta-analysis were indirect, the network was downgraded two levels.
Inconsistency	N/A: Inconsistency was marked as not applicable if there were no links in the network where data from multiple studies (either direct or indirect) were synthesised. For network meta-analyses conducted under a Bayesian framework, the network was downgraded one level if the DIC for an inconsistency model was more than 3 points lower than the corresponding consistency model or, for a

GRADE criteria	Reasons for downgrading quality
	random effects model, the between studies standard deviation was meaningfully lower for the inconsistency model than the corresponding consistency model.
Imprecision	This was not included in the GRADE table, but was considered during committee discussions of the evidence, taking into account 95% confidence intervals around the point estimate of the effect, any relevant MIDs, committee expertise and the effect of a single intervention based on multiple outcomes.

Diagnostic accuracy studies

Individual diagnostic accuracy studies were quality assessed using the QUADAS-2 tool. Each individual study was classified into one of the following three groups:

- Low risk of bias – The true effect size for the study is likely to be close to the estimated effect size.
- Moderate risk of bias – There is a possibility the true effect size for the study is substantially different to the estimated effect size.
- High risk of bias – It is likely the true effect size for the study is substantially different to the estimated effect size.

Each individual study was also classified into one of three groups for directness, based on if there were concerns about the population, index features and/or reference standard in the study and how directly these variables could address the specified review question. Studies were rated as follows:

- Direct – No important deviations from the protocol in population, index feature and/or reference standard.
- Partially indirect – Important deviations from the protocol in one of the population, index feature and/or reference standard.
- Indirect – Important deviations from the protocol in at least two of the population, index feature and/or reference standard.

GRADE for diagnostic accuracy evidence

Evidence from diagnostic accuracy studies was initially rated as high-quality, and then downgraded according to the modified GRADE criteria (risk of bias, inconsistency and indirectness) as detailed in Table 7 below.

The choice of primary outcome for decision making was determined by the committee and GRADE assessments were undertaken based on these outcomes.

In all cases, the downstream effects of diagnostic accuracy on patient- important outcomes were considered. This was done explicitly during committee deliberations and reported as part of the discussion section of the review detailing the likely consequences of true positive, true negative, false positive and false negative test results. In reviews where a decision model is being carried (for example, as part of an economic analysis), these consequences were incorporated here in addition.

Using sensitivity and specificity as the primary outcomes

GRADE assessments were only undertaken for sensitivity and specificity but results for positive and negative likelihood ratios are also presented alongside those data.

The committee were consulted to set 2 clinical decision thresholds for each measure: the value above which a test would be recommended, and a second below which a test would be considered of no clinical use. The committee decided that a sensitivity of 80% and specificity of 65% would be sufficient for a test to be considered as a potential diagnostic and monitoring tool for proliferative diabetic retinopathy or diabetic macular oedema.

If studies could not be pooled in a meta-analysis, GRADE assessments were undertaken for each study individually and reported as separate lines in the GRADE profile.

These criteria were used to apply preliminary ratings, but were overridden in cases where, in the view of the analyst or committee the uncertainty identified was unlikely to have a meaningful impact on decision making.

Table 6: Rationale for downgrading quality of evidence for diagnostic accuracy data

GRADE criteria	Reasons for downgrading quality
Risk of bias	Not serious: If less than 33.3% of the weight in a meta-analysis came from studies at moderate or high risk of bias, the overall outcome was not downgraded. Serious: If greater than 33.3% of the weight in a meta-analysis came from studies at moderate or high risk of bias, the outcome was downgraded one level. Very serious: If greater than 33.3% of the weight in a meta-analysis came from studies at high risk of bias, the outcome was downgraded two levels.
Indirectness	Not serious: If less than 33.3% of the weight in a meta-analysis came from partially indirect or indirect studies, the overall outcome was not downgraded. Serious: If greater than 33.3% of the weight in a meta-analysis came from partially indirect or indirect studies, the outcome was downgraded one level. Very serious: If greater than 33.3% of the weight in a meta-analysis came from indirect studies, the outcome was downgraded two levels.
Inconsistency	Concerns about inconsistency of effects across studies, occurring when there is unexplained variability in the treatment effect demonstrated across studies (heterogeneity), after appropriate pre-specified subgroup analyses have been conducted. This was assessed using the I^2 statistic. N/A: Inconsistency was marked as not applicable if data on the outcome was only available from one study. Not serious: If the I^2 was less than 33.3%, the outcome was not downgraded. Serious: If the I^2 was between 33.3% and 66.7%, the outcome was downgraded one level. Very serious: If the I^2 was greater than 66.7%, the outcome was downgraded two levels.
Imprecision	This was not included in the GRADE table, but was considered during committee discussions of the evidence, taking into account 95% confidence

GRADE criteria	Reasons for downgrading quality
	intervals around the point estimate of the effect, any relevant MIDs, committee expertise and the effect of a single intervention based on multiple outcomes.
Publication bias	If the review team became aware of evidence of publication bias (for example, evidence of unpublished trials where there was evidence that the effect estimate differed in published and unpublished data), the outcome was downgraded once. If no evidence of publication bias was found for any outcomes in a review (as was often the case), this domain was excluded from GRADE profiles to improve readability.

Association studies

Individual prognostic studies presenting data on association were quality assessed using the QUIPs checklist. Each study was classified into one of the following groups:

- Low risk of bias – The true effect size for the study is likely to be close to the estimated effect size.
- Moderate risk of bias – There is a possibility the true effect size for the study is substantially different to the estimated effect size.
- High risk of bias – It is likely the true effect size for the study is substantially different to the estimated effect size.

Each individual study was also classified into one of three groups for directness, based on if there were concerns about the population, factors and/or outcomes in the study and how directly these variables could address the specified review question. Studies were rated as follows:

- Direct – No important deviations from the protocol in population, factors and/or outcomes.
- Partially indirect – Important deviations from the protocol in one of the population, factors and/or outcomes.
- Indirect – Important deviations from the protocol in at least two of the population, factors and/or outcomes.

Modified GRADE for association data

GRADE has not been developed for use with association studies, therefore a modified approach was applied using the GRADE framework. Data from cohort, cross-sectional and case-control studies was initially rated as high quality, with the quality of the evidence for each outcome then downgraded or not from this initial point.

These criteria were used to apply preliminary ratings, but were overridden in cases where, in the view of the analyst or committee the uncertainty identified was unlikely to have a meaningful impact on decision making.

Table 7: Rationale for downgrading quality of evidence for association studies

GRADE criteria	Reasons for downgrading quality
Risk of bias	Not serious: If less than 33.3% of the weight in a meta-analysis came from studies at moderate or high risk of bias, the overall outcome was not downgraded. Serious: If greater than 33.3% of the weight in a meta-analysis came from studies at moderate or high risk of bias, the outcome was downgraded one level. Very serious: If greater than 33.3% of the weight in a meta-analysis came from studies at high risk of bias, the outcome was downgraded two levels. Outcomes meeting the criteria for downgrading above were not downgraded if there was evidence the effect size was not meaningfully different between studies at high and low risk of bias. In addition, unadjusted odds ratio outcomes from univariate analyses were downgraded one level, in addition to any downgrading for risk of bias in individual studies. Adjusted odds ratios from multivariate analyses were not similarly downgraded.
Indirectness	Not serious: If less than 33.3% of the weight in a meta-analysis came from partially indirect or indirect studies, the overall outcome was not downgraded. Serious: If greater than 33.3% of the weight in a meta-analysis came from partially indirect or indirect studies, the outcome was downgraded one level. Very serious: If greater than 33.3% of the weight in a meta-analysis came from indirect studies, the outcome was downgraded two levels. Outcomes meeting the criteria for downgrading above were not downgraded if there was evidence the effect size was not meaningfully different between direct and indirect studies.
Inconsistency	Concerns about inconsistency of effects across studies, occurring when there is unexplained variability in the treatment effect demonstrated across studies (heterogeneity). This was assessed using the I^2 statistic. N/A: Inconsistency was marked as not applicable if data on the outcome was only available from one study. Not serious: If the I^2 was less than 33.3%, the outcome was not downgraded. Serious: If the I^2 was between 33.3% and 66.7%, the outcome was downgraded one level. Very serious: If the I^2 was greater than 66.7%, the outcome was downgraded two levels. Outcomes meeting the criteria for downgrading above were not downgraded if there was evidence the effect size was not meaningfully different between studies with the smallest and largest effect sizes.
Imprecision	This was not included in the GRADE table, but was considered during committee discussions of the evidence, taking into account 95% confidence intervals around the point estimate of the effect, any relevant MIDs, committee expertise and the effect of a single intervention based on multiple outcomes.
Publication bias	If the review team became aware of evidence of publication bias (for example, evidence of unpublished trials where there was evidence that the effect estimate differed in published and unpublished data), the outcome was downgraded once. If no evidence of publication bias was found for any outcomes in a review (as was often the case), this domain was excluded from GRADE profiles to improve readability.

The quality of evidence for each outcome was upgraded if either of the following conditions were met:

- Data showing an effect size sufficiently large that it cannot be explained by confounding alone.
- Data where all plausible residual confounding is likely to increase our confidence in the effect estimate.

Reviewing economic evidence

Inclusion and exclusion of economic studies

Literature reviews seeking to identify published cost–utility analyses of relevance to the issues under consideration were conducted for all questions. In each case, the search undertaken for the clinical review was modified, retaining population and intervention descriptors, but removing any study-design filter and adding a filter designed to identify relevant health economic analyses. In assessing studies for inclusion, population, intervention and comparator, criteria were always identical to those used in the parallel clinical search; only cost–utility analyses were included. Economic evidence profiles, including critical appraisal according to the Guidelines manual, were completed for included studies.

Appraising the quality of economic evidence

Economic studies identified through a systematic search of the literature were appraised using a methodology checklist designed for economic evaluations (NICE guidelines manual; 2014). This checklist is not intended to judge the quality of a study per se, but to determine whether an existing economic evaluation is useful to inform the decision-making of the committee for a specific topic within the guideline.

There are 2 parts of the appraisal process. The first step is to assess applicability (that is, the relevance of the study to the specific guideline topic and the NICE reference case); evaluations are categorised according to the criteria in Table 12.

Table 8 Applicability criteria

Level	Explanation
Directly applicable	The study meets all applicability criteria, or fails to meet one or more applicability criteria but this is unlikely to change the conclusions about cost effectiveness
Partially applicable	The study fails to meet one or more applicability criteria, and this could change the conclusions about cost effectiveness
Not applicable	The study fails to meet one or more applicability criteria, and this is likely to change the conclusions about cost effectiveness. These studies are excluded from further consideration

In the second step, only those studies deemed directly or partially applicable are further assessed for limitations (that is, methodological quality); see categorisation criteria in Table 13.

Table 9 Methodological criteria

Level	Explanation
Minor limitations	Meets all quality criteria, or fails to meet one or more quality criteria but this is unlikely to change the conclusions about cost effectiveness
Potentially serious limitations	Fails to meet one or more quality criteria and this could change the conclusions about cost effectiveness
Very serious limitations	Fails to meet one or more quality criteria and this is highly likely to change the conclusions about cost effectiveness. Such studies should usually be excluded from further consideration

Where relevant, a summary of the main findings from the systematic search, review and appraisal of economic evidence is presented in an economic evidence profile alongside the clinical evidence.

Health economic modelling

As well as reviewing the published economic literature for each review question, as described above, original economic analysis was undertaken in selected areas. Priority areas for new health economic analysis were agreed by the committee.

The following general principles were adhered to in developing the analysis:

- Methods were consistent with the NICE reference case.
- The design of the model, selection of inputs and interpretation of the results was discussed and agreed with the committee.
- Where possible, model inputs were based on the systematic review of the public health literature, supplemented with other published data sources identified by the committee as required.
- When published data were not available committee expert opinion was used to populate the model.
- Model inputs and assumptions were reported fully and transparently.
- The results were subject to sensitivity analysis and limitations were discussed.

Full methods for the original cost-effectiveness analyses are described in the Health Economic Model Report.

References

Follmann D, Elliott P, Suh I, Cutler J (1992) Variance imputation for overviews of clinical trials with continuous response. Journal of Clinical Epidemiology 45:769–73

Fu R, Vandermeer BW, Shamliyan TA, et al. (2013) Handling Continuous Outcomes in Quantitative Synthesis In: Methods Guide for Effectiveness and Comparative Effectiveness Reviews [Internet]. Rockville (MD): Agency for Healthcare Research

and Quality (US); 2008-. Available from: http://www.ncbi.nlm.nih.gov/books/NBK154408/

Norman G., Sloan JA., Wyrwich KW. (2003) Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. Med Care 41(5):582-92.