National Institute for Health and Care Excellence

Final

Overweight and obesity management: preventing, assessing and managing overweight and obesity

NICE guideline NGXX

Methods

December 2024

NICE guideline: methods

Final

Evidence reviews were developed by the Guideline Development Team

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 All NICE guidance is subject to regular review and may be updated or withdrawn.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:

Contents

Development of the guideline

Remit

To see "What this guideline covers" and "What this guideline does not cover" please see the guideline scope (Weight management: preventing, assessing and managing overweight and obesity).

Methods

This guideline was developed in accordance with the process set out in 'Developing NICE guidelines: the manual (2022)'. Where the guidelines manual does not provide advice, additional methods are described below.

Methods specific to each review are described in the methods section of that review.

Developing the review questions and outcomes

The 6 review questions were developed for this guideline as part of this update. Three review questions were published part of previous updates. These review questions were based on the key areas identified in the guideline scope. Review questions were developed by the NICE Guideline Development Team (GDT) and refined, validated and signed off by the Guideline Committee and NICE quality assurance team.

The review questions were based on the PICO[S] framework - Population, Intervention, Comparator and Outcome [and Study type].

Full literature searches, critical appraisals and evidence reviews were completed for all review questions. Details of these elements are found in the review protocols for each review (see Appendix A of each relevant review). Where protocol deviations have been made, these will be reported in the 'Methods' section of the individual review.

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

Evidence review	Review questions	Type of review
A*	1.2. What are the most accurate and suitable anthropometric methods and associated boundary values for different ethnicities, to assess the health risk associated with overweight and obesity in adults, particularly those in black, Asian and minority ethnic groups?	Diagnostic and prognostic accuracy review
B*	1.1 What are the most accurate and suitable anthropometric methods and associated boundary values for different ethnicities, to assess the health risk associated with overweight, and obesity in children and young people, particularly those in black, Asian and minority ethnic groups?	Diagnostic and prognostic accuracy review
C*	2.2 What referral criteria for bariatric surgery are most effective to achieve weight loss and	Effectiveness review

Evidence review	Review questions	Type of review
	maintain a healthier weight in adults living with obesity?	
D	1.3a What approaches are effective and cost-effective in identifying overweight and obesity in children and young people, particularly those in black, Asian and minority ethnic groups? 1.3b What are the barriers and facilitators to identifying overweight and obesity in children and young people, particularly those in black, Asian and minority ethnic groups?	Mixed methods review
E	1.4a What approaches are effective and cost-effective in increasing uptake of weight management services in children and young people, particularly those in black, Asian and minority ethnic groups? 1.4b What are the barriers and facilitators to increasing uptake of weight management services in children and young people, particularly those in black, Asian and minority ethnic groups?	Mixed methods review
F	2.1 What is the effectiveness and cost effectiveness of total or partial meal replacements, intermittent fasting, plant-based and low carbohydrate diets, in achieving and maintaining weight loss in adults living with overweight or obesity?	Effectiveness review
G	2.3 What intervention components and approaches are effective, cost effective and acceptable for children and young people living with overweight or obesity?	Effectiveness review and qualitative review
H	2.4 What is the effectiveness and cost effectiveness of healthy living programmes for preventing overweight or obesity in children and young people?	Effectiveness review
I	2.5 What is the effectiveness, cost effectiveness and acceptability of psychological approaches to address the counterproductive effect of weight stigma in achieving or maintaining weight loss, or negating the adverse impact of weight stigma, in children, young people and adults?	Effectiveness review and qualitative review
* Reviews that were published as part of previous updates of guidelines. Methods specific to these reviews are included in the individual evidence reviews.		

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. Protocols are reproduced in each evidence review along with the PROPSERO registration number.

Searching for evidence

Evidence was searched for each review question using the methods specified in the NICE methods manual. Full details of search strategies, databases searched, and numbers of studies identified can be found in the appendices of each individual review.

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.

None of the evidence reviews made use of the priority screening functionality within the EPPI-reviewer software.

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 into the EPPI reviewer software.

Incorporating published evidence syntheses

For all review questions where a literature search was undertaken looking for a particular study design, published evidence syntheses (quantitative systematic reviews or qualitative evidence syntheses) containing studies of that design were also included. All included studies from those syntheses were screened to identify any additional relevant primary studies not found as part of the initial search. Evidence syntheses that were used solely as a source of primary studies were not formally included in the evidence review (as they did not provide additional data) and were not quality assessed.

Methods for combining evidence

Data synthesis for intervention studies

Where possible, meta-analyses were conducted to combine the results of quantitative studies for each outcome. Outcomes were deemed favourable if line of no effect was crossed.

Network meta-analyses was considered in situations where the following criteria were met:

- at least 3 treatment alternatives

- The aim of the review was to produce recommendations on the most effective option, rather than simply describe the effectiveness of treatment alternatives.

When there were 2 treatment alternatives, pairwise meta-analysis was used to compare interventions.

Where sufficient studies were available, meta-regression was considered to explore the effect of study level covariates. The meta-regression model estimates not only treatment effects but also the effect of each level of the covariate: e.g. the effect associated with delivering a treatment to the child alone, or to the parent alone, or to both the parent and child. Meta-regression models were run for each covariate separately.

Pairwise meta-analysis

Pairwise meta-analyses were performed in Cochrane Review Manager V5.3 where possible. 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). Where there were zero events in both intervention and control arms, studies were excluded from meta-analysis. Where there were zero events in one arm no adjustment was made (zero cell adjustment is not required with the Mantel–Haenszel method unless all studies have zero events in one arm in which case 0.5 is added to the arm).

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. Where outcomes measured the same underlying construct but used different instruments/metrics, data were analysed using standardised mean differences (SMDs, Hedges' g).

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. If only a subset of trials reported change from baseline data, final timepoint values were combined with change from baseline values to produce summary estimates of effect. For continuous outcomes analysed as standardised mean differences this was not possible. In this case, if all studies reported final timepoint data, this was used in the analysis. If some studies only reported data as a change from baseline, analysis was done on these data, and for studies where only baseline and final time point values were available, change from baseline standard deviations were estimated, assuming a correlation coefficient derived from studies reporting both baseline and endpoint data, or if no such studies were available, assuming a correlation of 0.5 as a conservative estimate (Follman et al., 1992; Fu et al., 2013). In cases where SMDs were used they were back converted to a single scale to aid interpretation by the committee where possible.

Network meta-analysis

Hierarchical Bayesian Network Meta-Analysis (NMA) was performed 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). The WinBUGS code provided in the appendices of the TSDs was used without substantive alteration to specify synthesis models.

In all models, results were assessed for convergence to determine the length of 'burn in' period required by examining the 'bgdiag' and 'history' plots. Additionally, 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 the case.

Change in BMI z-score:

Under 6 years old and 12-18 year olds

Two separate chains with different initial values were used. Results were reported summarising 60,000 samples from the posterior distribution of each model, having run and discarded the 'burn-in' iterations.

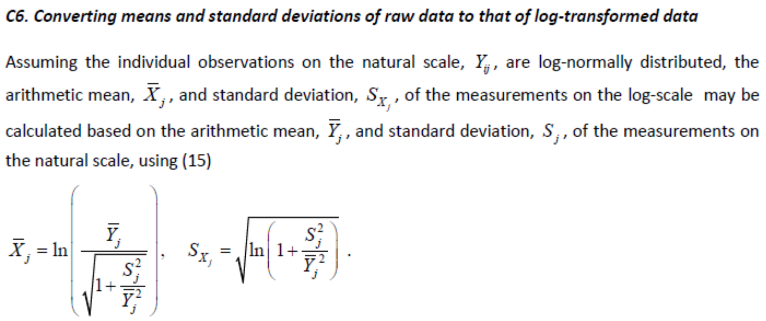
Under 6 -11 years old

Two separate chains with different initial values were used. Results were reported summarising 40,000 samples from the posterior distribution of each model, having run and discarded the 'burn-in' iterations.

Non-informative prior distributions

Non-informative prior distributions were used in under 6 years old, 6-11 years old (6-12 months follow up) and 12 -18 years old models. Unless otherwise specified, trial-specific baselines and treatment effects were assigned Normal (0, 10000) priors, and the between-trial standard deviations used in random-effects models for dichotomous outcomes were given Uniform (0, 5) priors. These are consistent with the recommendations in TSD 2 for dichotomous outcomes.

In the 6-11 years old, ≥ 12 months follow up model, informative priors were used. These priors were obtained from the 6-11 years old; 6-12 months follow up data. The first step in obtaining the informative priors was to run the random effects model and retrieve the SD. Using the mean (0.06546) and SD (0.02528) on the natural scale, the mean and SD were calculated on the lognormal scale using the following formulae (methods of moments approach).



C6. Converting means and standard deviations of raw data to that of log-transformed data

Assuming the individual observations on the natural scale, Y_{ij}, are log-normally distributed, the arithmetic mean, $\bar{X}_j$, and standard deviation, S_{X_j}, of the measurements on the log-scale may be calculated based on the arithmetic mean, $\bar{Y}_j$, and standard deviation, S_j, of the measurements on the natural scale, using (15)

$$\bar{X}_j = \ln\left(\frac{\bar{Y}_j}{\sqrt{1+\frac{S_j^2}{\bar{Y}_j^2}}}\right), \quad S_{X_j} = \sqrt{\ln\left(1+\frac{S_j^2}{\bar{Y}_j^2}\right)}.$$

Lastly, the standard prior for SD [dunif(0,5)] was replaced with the informative prior, where inf.sd is S_{xj} from the formula above and inf.M is Xbar_j and proceed as normal where the (inf.M) and SD of the informative prior (inf.sd) are provided as data.

```
sd.prec <- pow(inf.sd,-2)                # precision of informative distribution
sd ~ dlnorm(inf.M, sd.prec)    # prior on between-trial variance
```

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): 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 unrelated mean effects (UME) models. The consistency assumption is relaxed in the UME models, meaning that these can be used to check for inconsistency. The model fit was assessed using the deviance information criteria. A reduction in DIC of 3 or more was taken as evidence of inconsistency.

To visually assess if specific data-points were contributing to inconsistency, the deviance for each data-point in the NMA model was plotted against the UME model in dev-dev plots. Where residual deviance was relatively high (>2) in the NMA (consistency) model and lower in the UME model, studies were checked against the publications for accuracy and to assess their similarity to other studies reporting the same outcome and intervention."

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 qualitative reviews

Where multiple qualitative studies were identified for a single question, information from the studies was combined using a thematic synthesis. The thematic synthesis was based partly on a priori categories describing phenomena the committee was interested in and partly on themes that emerged from the coding of the included studies. Papers were uploaded to NVivo 11 software where the relevant data from the papers were coded. Once all the included studies had been examined and coded, the resulting sets of codes were aggregated into themes and sub-themes. The aggregated themes were used to develop interpretive 'review findings' that were evaluated using CERQual. These review findings were reproduced in a summary of qualitative findings table along with example quotes and details of the CERQual assessment of each review finding.

Data synthesis for mixed methods reviews

Data synthesis for mixed methods reviews was carried out in accordance with the Joanna Briggs Institute manual for evidence synthesis (https://wiki.jbi.global/display/MANUAL) chapter 8. Synthesis followed a convergent segregated approach where independent synthesis of quantitative data and qualitative data was undertaken, followed by the integration of the two types of evidence.

The qualitative and quantitative reviews were presented separately in the reviews and an integration section was written that addressed the following questions:

- Are the results/findings from individual syntheses supportive or contradictory?
- Does the qualitative evidence explain why the intervention is/is not effective?
- Does the qualitative evidence explain differences in the direction and size of effect across the included quantitative studies?
- Which aspects of the quantitative evidence were/were not explored in the qualitative studies?
- Which aspects of the qualitative evidence were/were not tested in the quantitative studies?

Where appropriate, any data from quantitative and qualitative sections of the review were integrated into tables or logic models/conceptual frameworks to show possible interrelationships between them.

Appraising the quality of evidence

Intervention studies (relative effect estimates)

RCTs and cluster randomised trials were quality assessed using the Cochrane Risk of Bias Tools. Review question 2.3 utilised Cochrane risk of bias tool 1 as this review incorporated existing Cochrane reviews which had utilised this tool. In order to main consistency between studies identified through the Cochrane reviews and new studies identified through searching, a decision was made to use ROB Tool 1. All other reviews utilised Cochrane risk of bias Tool 2.

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.

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 decision thresholds

The Core Outcome Measures in Effectiveness Trials (COMET) database was searched to identify published minimal important difference thresholds relevant to this guideline that might aid the committee in identifying decision thresholds for the purpose of GRADE. 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, committee members were asked to prospectively specify any outcomes where they felt a consensus decision threshold could be defined from their experience.

Decision thresholds were used to assess imprecision using GRADE and aid interpretation of the size of effects for different outcomes.

The following published MIDs were identified and discussed with the committee, and it was agreed they would be used across all reviews:

- Change in weight – 5% (Jensen et al. 2013)
- Change in HbA1c – 0.5% or 5 mmol/mol (Danker et al. 2021 and Little 2013)

For continuous outcomes expressed as a mean difference where no other decision threshold was available, a decision threshold of 0.5 of the median standard deviations of the comparison group arms was used (Norman et al. 2003). For continuous outcomes expressed as a standardised mean difference where no other decision threshold was available, a decision threshold of 0.5 standard deviations was used. For relative risks and hazard ratios, where no other clinical decision threshold was available, a default clinical decision threshold for dichotomous outcomes of 0.8 to 1.25 was used. The committee assessed the effects of the intervention by noting whether the effect estimate and 95% confidence intervals all lay to one side of the line of no effect. They agreed that when discussing interventions through a population level lens, any definite effect is a meaningful effect since even a very small effect multiplied across a large population will make a meaningful difference.

GRADE for pairwise meta-analyses of interventional evidence

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.

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	Outcomes were downgraded once if the 95% confidence interval for the effect size, crossed one of the default (0.8 and 1.25) or calculated MIDs, and twice if the effect estimate crossed both default (0.8 and 1.25) or calculated MIDs .
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.

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.

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 higher than the corresponding consistency model. For component NMAs, DIC was compared to standard NMA model statistics.
Imprecision	95% Credible intervals were used to assess imprecision. Not serious: The data were sufficiently precise to allow the committee to draw conclusions from the results of the NMA.

GRADE criteria	Reasons for downgrading quality
	Serious: Imprecision had a moderate impact on the ability of the committee to draw conclusions from the results of the NMA. Very serious: Imprecision had a substantial impact on the committee to draw conclusions from the results of the NMA.

GRADE-CERQual for qualitative evidence synthesis findings

CERQual was used to assess the confidence we have in each of the review findings. Evidence from all qualitative study designs (interviews, focus groups etc.) was initially rated as high confidence and the confidence in the evidence for each theme was assessed from this initial point as detailed in Table 7 below. Confidence in each criterion was assessed as:

- No or very minor concerns
- Minor concerns
- Moderate concerns
- Serious concerns

And an overall confidence rating of High, Moderate, Low or Very Low was determined based on this.

Table 6: Overall confidence in qualitative outcome

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

Table 7: Rationale for downgrading confidence in evidence for qualitative questions

CERQual criteria	Reasons for downgrading confidence
Methodological limitations	One or more studies contribute data to each review finding in a qualitative evidence synthesis, and these data make up the body of data for a review finding. The methodological limitations of the body of data supporting a review finding are assessed as a whole to identify whether or not any methodological weaknesses within individual studies impact our confidence in a review finding. The methodological limitations for each review finding must be assessed separately since different studies contribute varying amounts of data to each review finding, and methodological quality issues may have varying impacts on different review findings.
Relevance	Relevance is the extent to which the body of data from the primary studies supporting a review finding is applicable to the context specified

CERQual criteria	Reasons for downgrading confidence
	in the review question. Relevance is the CERQual component that is anchored to the context specified in the review question. How the review question and objectives are expressed, how a priori subgroup analyses are specified and how theoretical considerations inform the review design are therefore critical to making an assessment of relevance when applying CERQual.
Coherence	The coherence of a review finding is an assessment of how clear and cogent the fit is between the data from the primary studies and a review finding that synthesises that data. It includes consideration of the general 'fit' of data and whether any discrepancies can be explained.
Adequacy of data	Adequacy of data is an overall determination of the degree of richness as well as the quantity of data supporting a review finding. • Richness of the data is the extent to which the information that the individual study authors have provided is detailed enough to allow the review author to interpret the meaning and context of what is being researched. • Quantity of data relates to the number of studies and participants that this data comes from.

Modified GRADE-CERQual for published qualitative evidence synthesis (QES)

Published qualitative evidence syntheses only enable an indirect view of the evidence, through the interpretation of the QES authors. As such, it is not possible to fully apply GRADE-CERQual to their findings using the standard set of criteria as this requires direct examination of the data. For this reason, a modified version of the standard GRADE-CERQual approach for QESs was used to assess the quality of evidence identified through published qualitative evidence syntheses.

This version was modified based on the principles of GRADE-CERQual insofar as they could be applied to the information available, using the guidance provided in the Implementation Science series on 'Applying GRADE-CERQual to Qualitative Evidence Synthesis Findings'. While most criteria for qualitative evidence still apply, it is important to adapt some of the criteria to take into consideration additional factors, such how the published QESs reported quality of the evidence and the relevance of the individual studies to the review question. As a result, the following was used when modifying the GRADE-CERQual framework to published qualitative evidence syntheses.

Commented [MY1]: Changes made: both paragraphs rewritten

1 **Table 8: Rationale for downgrading confidence in evidence for published qualitative syntheses.**

CERQual criteria	Official guidance on applying CERQual to a qualitative evidence synthesis conducted by another review team[1]	Applied method for downgrading confidence in this review
Methodological limitations	CERQual does not recommend the use of a specific assessment tool and those applying CERQual need to judge if the tool used in a synthesis was appropriate. Some syntheses may present only an overall 'methodological quality' score for each included study. In these cases, the limitations of this for the CERQual assessment need to be acknowledged. In addition, some assessment tools include items related to adequacy and relevance. In these cases, care should be taken not to downgrade findings twice for the same concerns	Methodological limitations of a finding were assessed using the quality rating given to the individual studies by the QES authors. These ratings were adjusted to account for the appropriateness of the quality assessment tool or concerns with its implementation. • If minor concerns were identified with the reported quality assessment in the QES (for example, quality is provided but justifications not given or quality is not assessed using a standard tool or method), evidence from this QES was downgraded by 1 level. • If moderate concerns with the reported quality assessment (for example, standard tool or method not used and justification for risk of bias not given), evidence from this QES was downgraded by 2 levels. • If serious concerns were identified with the reported quality assessment (for example, if quality was not considered), evidence from this QES was downgraded by 3 levels. An individual study's contribution to a finding could not be determined without seeing the underlying data, so this could not be taken into account when assessing the methodological limitations of a finding. Therefore ratings of methodological limitations were based on an assumption of all studies contributing equally. • No concerns: all or most studies contributing to the finding were of low risk of bias. • Minor concerns: most studies contributing to the finding were of moderate or high risk of bias, but at least 1 was of low risk of bias. • Moderate concerns: all studies contributing to the finding were of moderate or high risk of bias. • Serious concerns: all studies contributing to the finding were of high risk of bias.
Relevance	It is not possible to assess this component if a synthesis does not include a 'Characteristics of included studies' table as insufficient detail or missing characteristics may impair the quality of the relevance judgements. CERQual cannot be applied to such syntheses	Individual studies included in QES were assessed in terms of their relevance to the overall review question (not to the individual QES review questions), based on how well the population, setting, date, and intervention type match the review protocol. An individual study's contribution to a theme could not be determined without seeing the underlying data, so this could not be taken into account when assessing the relevance of the theme. Therefore ratings of relevance were based on an assumption of all studies contributing equally.

CERQual criteria	**Official guidance on applying CERQual to a qualitative evidence synthesis conducted by another review team**[1]	**Applied method for downgrading confidence in this review**
	without going back to the included primary studies	• No concerns: all or most studies contributing to the finding were assessed as relevant to the overall review question. • Minor concerns: most studies contributing to the finding were indirectly or partially relevant to the overall review question, but at least 1 was relevant. • Moderate concerns: all studies contributing to the finding were indirectly or partially relevant to the overall review question. • Serious concerns: all studies contributing to the finding were partially relevant to the overall review question.
Coherence	Unless a synthesis presents detailed tables of the data contributing to each review finding, it may not be possible to assess this component	Coherence was based on the rating of coherence as reported by the QES authors, where possible. If the authors did not report coherence, then the finding was rated having minor concerns by default, as it was not possible to assess coherence without seeing the underlying data. • No concerns: If the review authors report no concerns with coherence. • Minor concerns: If the review authors reported minor concerns with coherence or if the authors did not assess coherence. • Moderate concerns: If the review authors reported moderate concerns with coherence. • Serious concerns: If review authors reported serious concerns with coherence.
Adequacy of data	For many syntheses, adequacy may need to be assessed based on only the number of studies contributing data to a review finding and not the depth of the data. This limitation needs to be acknowledged.	Adequacy of data was based primarily on the number of studies included in the analysis, as it was not generally possible to assess the depth or richness of the data without seeing the underlying data. • No concerns: findings came from 3 studies or more. • Minor concerns: findings came from fewer than 3 studies • Moderate concerns: findings came from a single study, but the descriptions suggested richness or multifaceted data. • Serious concerns: findings came from a single study and richness could not be determined

Commented [MY2]: Changes made: Official guidance column added and applied method column reworded for clarity.

[1] Lewin, S., Bohren, M., Rashidian, A. *et al.* Applying GRADE-CERQual to qualitative evidence synthesis findings—paper 2: how to make an overall CERQual assessment of confidence and create a Summary of Qualitative Findings table. *Implementation Sci* **13** (Suppl 1), 10 (2018). https://doi.org/10.1186/s13012-017-0689-2

Mixed methods studies

Mixed methods studies were evaluated using the appropriate quality assessment tools for the component study types, see sections on intervention studies and qualitative studies. Other methods of assessing mixed methods studies were agreed with the NICE methods and economics team QA lead and reported in the individual reviews.

Using Cochrane reviews

During the development of the review question 2.3, a Cochrane review (Brown 2019) was identified. This review was due to be updated and protocols covering children and young people aged 2-18 years old were identified as being relevant to the NICE review. These updates were identified as being directly applicable to the review question, based on the criteria outlined in Table 2.

The evidence presented in RQ2.3 is based on the Cochrane reviews produced by University of Bristol and University of Durham as part of a collaboration between the NICE Guideline Development Team and Cochrane.

As part of the collaboration, authors based at Bristol University and Durham University preformed:

- The literature search, screening of records and study selection.
- Data extraction and production of evidence tables.
- Risk of bias assessment of included studies using Cochrane Risk of bias tool 2
- Publication bias assessment using funnel plots.
- Data analysis, including pairwise meta-analysis, subgroup analysis by setting.
- Presentation of evidence to guideline committee.

The NICE Development Team assisted in the data extraction, risk of bias and data analysis included in the review covering children aged 2-4 years old.

Approaches used to search the evidence, selecting studies for inclusion and synthesising the evidence are detailed in the Cochrane reviews.

Commented [SS3]: We will need to add the links to the reviews.

GRADE approach utilised in Cochrane reviews

GRADE approach utilised in the Cochrane reviews differed to that detailed in section GRADE for pairwise meta-analyses for intervention evidence. Table 9 details the approaches utilised by the authors.

Table 9: Rationale for downgrading confidence in evidence used in Cochrane review.

GRADE criteria	Reasons for downgrading confidence
Risk of bias	Based on results of our risk of bias assessments, we downgraded confidence in the evidence base if most evidence was from studies that we judged at high risk of bias, according to the following rules: • No serious concerns (no downgrade): contributing weight of evidence at high risk < 30%. • Serious concerns (one point down): contributing weight of evidence of high risk of bias > 30%. • Very serious concerns (two points down): not applied.

GRADE criteria	Reasons for downgrading confidence
Imprecision	We downgraded confidence in the evidence base if the estimate of the effect size from a meta- analysis was not precise, according to the following rules: • -No serious concerns (no downgrade): >3000 participants or clear evidence of an effect larger than ± 1/5 of a typical standard deviation (which corresponds to 0.2 for zBMI, 0.5 for BMI or 6 for BMI percentile). • -Serious concerns (one point down): <3000 participants without clear evidence of an effect larger than ± 1/5 of a typical standard deviation. • -Very serious concerns (two points down): not applied.
Inconsistency	We downgraded confidence in the evidence base if there was unexplained heterogeneity or variability in results across studies, according to the following rules: • No serious concerns (no downgrade): estimated heterogeneity variance (tau) = 0 or results all in the same direction. • Possible serious concern (half point down): estimated heterogeneity variance (tau) of moderate magnitude and the direction of the results is inconsistent; or the results are from a single study. • Serious concerns (one point down): estimated heterogeneity variance (tau) is high and the direction of the results is inconsistent. • Very serious concerns (two points down): not applied.
Indirectness	We downgraded confidence in the evidence base if we had concerns that the population was highly specific and reducing the generalisability of the results, according to the following rules: • No serious concerns (no downgrade): no study populations of concern, or contributing weight of studies in highly specific populations <30%. • Serious concerns (one point down): contributing weight of studies in highly specific populations >30%. • Very serious concerns (two points down): not applied.
Non-reporting bias	We downgraded our confidence in the evidence base due to within-study non-reporting if there was (i) evidence of outcome measurement and (ii) indication of unreported non-statistically-significant result(s) and (iii) potential for the missing result(s) to impact on the meta-analysis, according to the following rules: • No serious concerns (no downgrade): no missing outcome data, or studies with missing outcome data were not large enough to impact on meta-analyses. • Possible serious concern (half point down): we had evidence of measured outcomes being missing but no indication of the reason, and missing studies were potentially large enough to affect the result. • Serious concerns (one point down): we had evidence of measured outcomes being missing and an indication that missing results were not statistically significant and able to affect the meta-analyses result. • Very serious concerns (two points down): not applied. • We considered that any wholly missing studies were likely to be small, whereas many included studies are large. We therefore did not have strong reason to rate down for publication bias in addition to selective non-reporting within studies.

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 public health 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 public health 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; 2020). 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 10.

Table 10 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 11.

Table 11 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 public health 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.

Resource impact assessment

The resource impact team used the methods outlined in the in Assessing resource impact process manual: guidelines

The resource impact team worked with the guideline committee from an early stage to identify recommendations that either individually or cumulatively would a substantial impact on resources. The aim was to ensure that a recommendation would not introduce a cost pressure into the health and social care system unless the committee was convinced of the benefits and cost effectiveness of the recommendation. The team gave advice to the committee on issues related to the workforce, capacity and demand, training, facilities and educational implications of the recommendations.

Peer review of the draft guideline

The draft guideline was peer reviewed by Jamie Blackshaw, Office for Health Improvement and Disparities, prior to draft guideline stakeholder consultation. This was carried out to check the policy context and referencing to statutory guidance and policy documents.