National Institute for Health and Care Excellence

Guideline version (for example, Draft, Consultation, Final)

Looked-After Children and Young People (update)

NICE guideline: methods

NICE guideline <NGXXX>

Appendix N

[April 2021]

Draft for Consultation

Evidence reviews were developed by the NICE guideline updates 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 All NICE guidance is subject to regular review and may be updated or withdrawn.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

This guideline will update and replace the NICE guideline on looked-after children and young people (PH28).

This guideline will also be used to update the NICE quality standard for looked-after children and young people.

To see "What this guideline covers" and "What this guideline does not cover" please see the guideline scope for Looked-after children and young people

Methods

This guideline was developed using the methods described in the 2018 NICE guidelines manual.
Declarations of interest were recorded according to the NICE conflicts of interest policy.

Developing the review questions and outcomes

The 15 review questions developed for this guideline were based on the key areas identified in the guideline scope. They were drafted by the NICE guideline updates team, and refined and validated by the guideline committee.
The review questions were based on the following frameworks:

- population, intervention, comparator and outcome (PICO) for reviews of interventions
- sample, phenomenon of interest, design, evaluation, (SPiDEr) for qualitative review questions

Full literature searches, critical appraisals and 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 2018 NICE guidelines manual.

Selecting studies for inclusion

All references identified by the searches and from other sources (for example, a previous version of the guideline or studies identified by committee members) were uploaded into EPPI reviewer software 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 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. Study investigators were contacted for missing data when time and resources allowed.

Incorporating published evidence syntheses

For all review questions where a literature search was undertaken looking for a particular study design, 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. Systematic reviews 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. Committee members were also consulted to identify studies that may have been missed.

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 the following criteria were met:

- At least three 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.

In other situations, pairwise meta-analysis was used to compare interventions.

Pairwise meta-analysis

Pairwise meta-analyses were performed in Cochrane Review Manager V5.3, with the exception of incidence rate ratio analyses which were carried out in R version 3.3.4. using the package ‘metafor’. A pooled relative risk was calculated for dichotomous outcomes (using the Mantel–Haenszel method) reporting numbers of people having an event, and a pooled incidence rate ratio was calculated for dichotomous outcomes reporting total numbers of events. 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. Where outcomes measured the same underlying construct but used different instruments/metrics, data were analysed using standardised mean differences (Hedges' g).

For continuous outcomes analysed as mean differences, where change from baseline data were reported in the trials and were accompanied by a measure of spread (for example standard deviation), these were extracted and used in the meta-analysis. Where measures of spread for change from baseline values were not reported, the corresponding values at study end were used and were combined with change from baseline values to produce summary estimates of effect. These studies were assessed to ensure that baseline values were balanced across the treatment groups; if there were significant differences at baseline these studies were not included in any meta-analysis and were reported separately. For continuous outcomes analysed as standardised mean differences, where only baseline and final time point values were available, change from baseline standard deviations were estimated, assuming a correlation coefficient of 0.5. In cases where SMDs were used they were back converted to a single scale to aid interpretation by the committee where possible.

Fixed- and random-effects models were fitted for all syntheses, 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 one or both of the following conditions was met:

- 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.
- The presence of 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.

In any meta-analyses where some (but not all) of the data came from studies at high or critical risk of bias, a sensitivity analysis was conducted, excluding those studies from the analysis. Similarly, in any meta-analyses where some (but not all) of the data came from indirect studies, a sensitivity analysis was conducted, excluding those studies from the analysis. Results from both the full and restricted meta-analyses are reported.

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.

At least two separate chains with different initial values were used. Results were assessed for convergence to determine the length of 'burn in' period required by examining the 'bgdiag' and 'history' plots. Results were reported summarising at least 10,000 samples from the posterior distribution of each model, having run and discarded the 'burn-in' iterations. 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.

Non-informative prior distributions were used in all 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.

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.

In any meta-analyses where some (but not all) of the data came from studies at high risk of bias, a sensitivity analysis was conducted, excluding those studies from the analysis. Results from both the full and restricted meta-analyses are reported. Similarly, in any meta-analyses where some (but not all) of the data came from indirect studies, a sensitivity analysis was conducted, excluding those studies from the analysis. Where sufficient studies were available, meta-regression was considered to explore the effect of study level covariates.

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 inconsistency.

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. Papers were uploaded to NVivo 11 software where the relevant themes from the papers were coded. Once all of the included studies had been examined and coded, the resulting aggregated themes and sub-themes were evaluated to examine their relevance to the review question, the importance given to each theme, and the extent to which each theme recurred across

the different studies. The aggregated themes were used to develop interpretive 'review findings'. 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.

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). 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 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, the Guideline Committee were asked to prospectively specify any outcomes where they felt a consensus decision threshold could be defined from their experience. However, this option was not used by the Guideline Committee for Looked After Children and Young People for any identified outcome.

Therefore, 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 was used. For relative risks where no other decision threshold was available, a default decision threshold for dichotomous outcomes of 0.8 to 1.25 was used.

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 1.

Table 1: 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 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.
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), 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. 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	If an MID other than the line of no effect was defined for the outcome, the outcome was downgraded once if the 95% confidence interval for the effect size crossed one line of the MID, and twice if it crosses both lines of the MID.

GRADE criteria	Reasons for downgrading quality
	If the line of no effect was defined as an MID for the outcome, it was downgraded once if the 95% confidence interval for the effect size crossed the line of no effect (i.e. the outcome was not statistically significant), and twice if the sample size of the study was sufficiently small that it is not plausible any realistic effect size could have been detected. Outcomes meeting the criteria for downgrading above were not downgraded if the confidence interval was sufficiently narrow that the upper and lower bounds would correspond to equivalent scenarios.
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 studies that were originally assigned a quality rating of 'low' (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:

- 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.

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, which can then be combined with pairwise quality ratings for individual comparisons (if appropriate), to judge the overall strength of evidence for each comparison.

Table 2: 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.

GRADE criteria	Reasons for downgrading quality
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.
Imprecision	Not serious: The data were sufficiently precise to meet the aims of the review question. Serious: Imprecision had a moderate impact on the ability of the data to meet the aims of the review question. Very serious: Imprecision had a substantial impact on the ability of the data to meet the aims of the review question.

Qualitative studies

Individual qualitative studies were quality assessed using the CASP qualitative checklist. Each individual study was classified into one of the following three groups:

- Low risk of bias – The findings and themes identified in the study are likely to accurately capture the true picture.
- Moderate risk of bias – There is a possibility the findings and themes identified in the study are not a complete representation of the true picture.
- High risk of bias – It is likely the findings and themes identified in the study are not a complete representation of the true picture

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

- Highly relevant – No important deviations from the protocol in perspective, population, phenomenon of interest and/or setting.
- Relevant – Important deviations from the protocol in one of the perspective, population, phenomenon of interest and/or setting.
- Partially relevant – Important deviations from the protocol in at least two of the perspective, population, phenomenon of interest and/or setting.

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 then downgraded from this initial point as detailed in Table 3 below.

Table 3 Rationale for downgrading confidence in evidence for qualitative questions

CERQual criteria	Reasons for downgrading confidence
Methodological limitations	Not serious: If the theme was identified in studies at low risk of bias, the outcome was not downgraded Serious: If the theme was identified only in studies at moderate or high risk of bias, the outcome was downgraded one level. Very serious: If the theme was identified only in studies at high risk of bias, the outcome was downgraded two levels.

CERQual criteria	Reasons for downgrading confidence
Relevance	High: If the theme was identified in highly relevant studies, the outcome was not downgraded Moderate: If the theme was identified only in majority partially relevant studies, the outcome was downgraded one level. Low: If the theme was identified only in partially relevant studies, the outcome was downgraded two levels.
Coherence	Coherence was addressed based on two factors: Between study – does the theme consistently emerge from all relevant studies Theoretical – does the theme provide a convincing theoretical explanation for the patterns found in the data The outcome was downgraded once if there were concerns about one of these elements of coherence, and twice if there were concerns about both elements.
Adequacy of data	The outcome was downgraded if there was insufficient data to develop an understanding of the phenomenon of interest, either due to insufficient studies, participants or observations.

Health economics

No de novo economic models were built for this guideline. However, a costing analysis was conducted to support a recommendation made for review questions 2.1 and 3.2. Further details outlining the rationales for not building any de novo economic models for this guideline and the methods used to undertake the costing analysis are provided in the evidence reviews for review questions 2.1 and 3.2. Literature reviews seeking to identify published cost–effectiveness and cost-utility analyses of relevance to the issues under consideration were conducted for all questions. In each case, the search undertaken for the 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 search; only cost–effectiveness and cost-utility analyses were included. Economic evidence profiles, including critical appraisal according to the Guidelines manual, were completed for included studies.

Economic studies identified through a systematic search of the literature are 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 9.

Table 9 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

Level	Explanation
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 9 Applicability **criteria**

Table 10 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 review evidence.