The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions

Sara H Downs, Nick Black

Abstract

***Objective*—To test the feasibility of creating a valid and reliable checklist with the following features: appropriate for assessing both randomised and non-randomised studies; provision of both an overall score for study quality and a profile of scores not only for the quality of reporting, internal validity (bias and confounding) and power, but also for external validity.**

***Design*—A pilot version was first developed, based on epidemiological principles, reviews, and existing checklists for randomised studies. Face and content validity were assessed by three experienced reviewers and reliability was determined using two raters assessing 10 randomised and 10 non-randomised studies. Using different raters, the checklist was revised and tested for internal consistency (Kuder-Richardson 20), test-retest and inter-rater reliability (Spearman correlation coefficient and sign rank test; κ statistics), criterion validity, and respondent burden.**

***Main results*—The performance of the checklist improved considerably after revision of a pilot version. The Quality Index had high internal consistency (KR-20: 0.89) as did the subscales apart from external validity (KR-20: 0.54). Test-retest (*r* 0.88) and inter-rater (*r* 0.75) reliability of the Quality Index were good. Reliability of the subscales varied from good (bias) to poor (external validity). The Quality Index correlated highly with an existing, established instrument for assessing randomised studies (*r* 0.90). There was little difference between its performance with non-randomised and with randomised studies. Raters took about 20 minutes to assess each paper (range 10 to 45 minutes).**

***Conclusions*—This study has shown that it is feasible to develop a checklist that can be used to assess the methodological quality not only of randomised controlled trials but also non-randomised studies. It has also shown that it is possible to produce a checklist that provides a profile of the paper, alerting reviewers to its particular methodological strengths and weaknesses. Further work is required to improve the checklist and the training of raters in the assessment of external validity.**

(*J Epidemiol Community Health* 1998;**52**:377–384)

Health Services Research Unit, Department of Public Health and Policy, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT

Correspondence to: Professor Black.

Accepted for publication 28 August 1997

Clinicians are increasingly being encouraged to base their decisions on scientific evidence from systematic reviews and meta-analyses.[1] This is reflected in the establishment of organisations whose primary function is to conduct reviews, such as the Cochrane Collaboration and, in the UK, the NHS Centre for Reviews and Dissemination. Systematic reviews must seek to be comprehensive in terms of the evidence they examine and objective in the way in which the evidence is judged. Generally, such reviews have concentrated exclusively on randomised trials. Indeed, some investigators are of the opinion that non-randomised (or observational) studies should be excluded from all reviews because of the greater difficulties in assessing their methodological quality. However, in many areas of health care few randomised controlled trials exist and most of those that have been done have been poorly executed.[2 3]

Whereas at least 25 checklists have been developed that provide a framework for judging the methodological quality of randomised trials,[4] a Medline search from 1990 to January 1997 failed to identify any for the assessment of analytical non-randomised studies (cohort and case-control studies).

Although the design of randomised trials, cohort studies, and case-control studies have fundamental differences, in all designs three factors are measured: the intervention, potential confounders, and the outcome. All test to see if there is an association between the intervention and the outcome and aim to minimise flaws in the design that will bias the measurement of an association. The vulnerability of each design to different biases varies but the kind of biases that the study designs seek to exclude are the same.

A second limitation of existing instruments is their lack of sub-scales that provide a profile of the strengths and weaknesses of each methodological concern. This is important because the particular defects of a study determine how it may be interpreted. For example a reader informed that the design has not tackled selection bias may consider whether, in this instance, confounding is likely to be a major problem. Or if the main problem of the study was inadequate power, the reader might choose to place less weight on a null finding.

The third limitation of existing instruments is their exclusion of any consideration of external validity (generalisability). No explanation is offered as to why this methodological aspect is

ignored, beyond a belief that it is of little importance.[5 6]

The objective of this study was to test the feasibility of creating a valid and reliable checklist with the following features: appropriate for assessing both randomised and non-randomised studies; providing both an overall score for study quality and a profile of scores not only for the quality of reporting, internal validity (bias and confounding) and power, but also for external validity.

Methods

DEVELOPMENT OF A PILOT VERSION

A pilot version of the checklist was developed based on epidemiological principles, reviews of study designs,[7–9] and existing checklists for the assessment of randomised controlled trials.[4 10–15] The pilot checklist consisted of 26 items distributed between five sub-scales:

(1) Reporting (9 items)—which assessed whether the information provided in the paper was sufficient to allow a reader to make an unbiased assessment of the findings of the study.

(2) External validity (3 items)—which addressed the extent to which the findings from the study could be generalised to the population from which the study subjects were derived.

(3) Bias (7 items)—which addressed biases in the measurement of the intervention and the outcome.

(4) Confounding (6 items)—which addressed bias in the selection of study subjects.

(5) Power (1 item)—which attempted to assess whether the negative findings from a study could be due to chance.

Answers were scored 0 or 1, except for one item in the Reporting subscale, which scored 0 to 2 and the single item on power, which was scored 0 to 5. The total maximum score was therefore 31. Most (23) of the questions could be asked of any analytical study of any health care intervention. Three questions, however, were inevitably topic sensitive and had to be customised by providing the raters with information on: known confounders; main outcomes; and the sample size required for a clinically and statistically significant ($p<0.05$) result.

TESTING OF PILOT VERSION

Two senior epidemiologists and a medical statistician were asked to comment on the face and content validity of the checklist after which some modifications were made. Two raters, both of whom were non-medical research fellows with Masters degrees in epidemiology and who had not been involved in the development of the checklist, were asked to assess 10 randomised controlled trials and 10 non-randomised trials/prospective cohort studies selected at random from a group of 11 randomised controlled trials and 20 cohort studies identified during a systematic review of surgery for stress incontinence.[3] Authors' identity, institutional affiliations, and journal identity were removed. Four of the papers were translations from German and one from Spanish. The raters were given guidance with regard to the interpretation of items included in the checklist before reviewing the papers.

Inter-rater reliability was assessed as was test-retest reliability, by both raters repeating the exercise after two weeks. Criterion validity was assessed by comparing the Quality Index (total score) with the total score obtained using an existing validated checklist,[15] though inevitably this was restricted to the randomised controlled trials. The level of association was assessed by means of Spearman correlation coefficients and the level of agreement by means of the κ statistic.[16]

Inter-rater reliability (including all 26 items) was only modest (correlation coefficient, r=0.47; κ 0.42) and this was true both for randomised controlled trials (r 0.43; κ 0.39) and non-randomised studies (r 0.50; κ 0.46). Some reasons for this became apparent both by considering individual items and during feedback on the face validity of the checklist. Test-retest reliability was fair for both raters (rater A: r 0.68, κ 0.66; rater B: r 0.64, κ 0.64) as was the criterion validity (r 0.78, κ 0.61).

DEVELOPMENT OF REVISED VERSION

Given the less than satisfactory psychometric properties of the pilot version, a new version was produced (appendix). Any items for which the answers would be the same for most papers were removed as such items contribute nothing to the discriminant properties of the checklist. All items were kept as short as possible.

The revised version incorporated an extra item in the Reporting sub-scale. In addition, a global item was included at the end of the checklist, which asked the raters to give the paper a Global Score out of 10, to see how the Quality Index score compared with the raters' overall impression of the quality of a paper.

TESTING OF THE REVISED VERSION

A new rater, a non-medical research fellow with a Masters degree in epidemiology, who had not been involved in the development of the checklist was engaged to review the same selection of papers. The rater was given guidance with regard to the interpretation of questions before reviewing the papers. The following psychometric properties of the checklist were assessed[17]:

(1) Internal consistency was tested using the Kuder-Richardson formula 20 (KR-20) as all but two items employed a dichotomous response.[18] The internal consistency of the rater's assessments was studied both for all the papers and for randomised and non-randomised studies separately. The internal consistency of four of the five sub-scales (power was based on only one item) was also assessed.

(2) Test-retest reliability was assessed by asking the rater to repeat her assessment of each paper after a two week interval.[19] Association and agreement of the Quality Index scores, sub-scales, and individual items were investigated using Spearman correlation coefficients (as the data were not normally

Table 1 Crude summary data on quality of 20 papers assessed using the checklist

	Randomised studies		*Non-randomised studies*	
	Mean	*Range*	*Mean*	*Range*
Reporting	5.9	2–10	6.1	0–10
External validity	0.3	0–2	0.1	0–1
Bias	4.2	0–6	4.0	1–5
Confounding	3.6	1–6	1.5	0–3
Power	0	0	0	0

Table 2 Internal consistency of Quality Index and sub-scales (KR-20)

Scale (items)	*RCT + non-randomised*	*RCT*	*Non-randomised*
Quality Index (26)	0.89	0.92	0.88
Reporting (10)	0.79	0.80	0.83
Confounding (6)	0.69	0.74	0.48
Bias (7)	0.78	0.86	0.78
External validity (3)	0.54	0.68	0.15
Internal + external validity (16)	0.72	0.81	0.65

RCT=randomised controlled trial.

Table 3 Test-retest reliability (Spearman correlation coefficients)

	RCT + non-randomised		*RCT*		*Non-randomised*	
Scale	*Correlation*	*p Value*	*Correlation*	*p Value*	*Correlation*	*p Value*
Quality Index	0.88	0.90	0.76	1.00	0.79	0.84
Report	0.84	0.68	0.88	0.54	0.73	0.92
Confounding	0.69	0.0008	0.77	0.10	0.53	0.31
Bias	0.90	0.13	0.85	0.11	0.86	0.65
External validity	0.37	0.33	0.25	0.96	0.65	0.17

p Value based on sign rank test.

Table 4 Test-retest reliability and inter-rater reliability of items (κ and percentage disagreement) based on 20 papers (RCTs and non-randomised studies)

		Test-retest reliability		*Inter-rater reliability*	
Subscale	*Item*	*κ*	*% disagree*	*κ*	*% disagree*
Reporting	1	0.63	15	0.00	25
	2	0.39	25	0.08	35
	3	0.80	10	0.48	25
	4	0.89	5	0.00	30
	5	0.69	10	0.38	20
	6	0.57	15	0.17	25
	7	0.63	15	0.55	20
	8	0.68	15	0.51	25
	9	−0.18	40	0.21	30
	10	0.88	5	0.88	5
External validity	11	0.48	15	−0.08	20
	12	−0.05	10	0.00	5
	13	0.00	15	0.00	5
Bias	14	0.22	25	0.12	30
	15	1.00	0	1.00	0
	16	0.38	30	0.00	35
	17	0.06	40	0.30	35
	18	0.35	30	0.22	20
	19	−0.05	10	0.00	5
	20	0.76	10	0.58	15
Confounding	21	0.88	5	0.78	10
	22	0.70	15	0.80	10
	23	1.00	0	1.00	0
	24	0.74	10	0.00	20
	25	0.78	10	0.47	20
	26	0.27	20	0.29	25

distributed) and κ statistics. To assess statistical significance, sign rank tests that took into account the paired nature of the data were used to compare the distribution of scores for the Quality Index and each sub-scale.

(3) Inter-rater reliability was assessed by comparing the primary rater's assessment with ratings obtained from a second rater (a medical graduate with a Masters degree in epidemiology). Association and agreement were tested for in the same way as for test-retest reliability.

(4) Criterion validity of the checklist when used with randomised controlled trials was assessed by comparing the total scores with those obtained using another checklist (the SRTG[15]) designed exclusively for randomised controlled trials, which comprised 32 items. Correlation of the Quality Index score with the Global Score provided another view of criterion validity (both for randomised controlled trials and cohort studies).

(5) Respondent burden was assessed in terms of the time required for assessing each paper and the raters' views of the level of knowledge required.

Results

CRUDE SUMMARY DATA

The mean (SD) Quality Index Score for randomised controlled trials was 14.0 ((6.39); skewness −0.07) and for non-randomised studies 11.7 ((SD) 4.64; skewness −1.10). Table 1 shows the mean scores and range of scores for each sub-scale.

INTERNAL CONSISTENCY (KR-20)

The internal consistency of the Quality Index was high (KR-20: 0.89) both for randomised and non-randomised studies (table 2). Three of the four sub-scales showed adequate internal consistency. The exception was external validity (KR-20: 0.54), which arose partly from the small number of items making up the sub-scale and partly because of its poor performance with non-randomised studies (KR-20: 0.15).

A sub-scale that combined internal (bias and confounding) and external validity was also investigated. This displayed reasonably high internal consistency both for randomised and non-randomised studies (KR-20 = 0.72).

TEST-RETEST RELIABILITY

The test-retest reliability of the Quality Index was high ($r = 0.88$) (table 3). The reliability of the sub-scales varied from high (bias) to low (external validity). The sign rank test showed no difference in the distributions of scores for the Quality Index and its sub-scales, except for "confounding" when both randomised controlled trials and cohort studies were tested together. To investigate the performance of the sub-scales, the level of agreement for each item was considered (table 4). Only one of the 10 items making up the Reporting sub-scale (item 9) showed poor agreement (κ = −0.18) and only one of the six items comprising the Confounding sub-scale (item 26; κ = 0.27). In contrast, three of the seven Bias items were poor (item 14, 0.22; item 17, 0.06; item 19, −0.05) as were two of the three items for External validity (item 12, −0.05; item 13, 0.00).

INTER-RATER RELIABILITY

The inter-rater reliability of the Quality Index was good ($r = 0.75$) (table 5). The reliability of the sub-scales varied from good (bias) to poor (external validity). The sign rank test showed no difference in the distributions of scores for the Quality Index and its sub-scales. There was

Table 5 Inter-rater reliability of the Quality Index, sub-scales, and the SRTG (Spearman correlation coefficients)

	RCT + non-randomised		*RCT*		*Non-randomised*	
Scale (items)	*Correlation*	*p Value*	*Correlation*	*p Value*	*Correlation*	*p Value*
Quality Index	0.75	0.56	0.73	0.44	0.77	1.00
Reporting (10)	0.71	0.09	0.78	0.08	0.51	0.51
Confounding (6)	0.34	0.14	−0.07	0.88	0.45	0.76
Bias (7)	0.83	0.34	0.78	0.72	0.59	0.33
External validity (3)	−0.14	0.71	−0.25	0.92	0.00	0.61
SRTG (32)	NA	NA	0.80	0.006	NA	NA

p Value based on sign rank test.

no difference in the distributions when randomised controlled trials and cohort studies were considered separately. To investigate the performance of the sub-scales, the level of agreement for each item was considered (table 3). All three items making up the External validity sub-scale contributed to its poor performance.

For comparative purposes, the inter-rater reliability of an established checklist[15] for assessing randomised controlled trials was also measured (table 5). The reliability of the overall score of the SRTG's instrument was similar to that obtained for the new Quality Index. The proportion of SRTG's items with poor agreement (κ scores of less than 0.2) was 41% (13 of 32) compared with 42% (11 of 26) for our new checklist.

CRITERION VALIDITY

The Quality Index score correlated highly (0.90) with the score obtained using the instrument of the SRTG (randomised controlled trials only) and with the Global Score (randomised controlled trials + non-randomised studies, 0.89; randomised controlled trials, 0.88; non-randomised studies, 0.86). It should be noted, however, that the reviewer assigned a global score after completing the checklist of 27 items so a high correlation would be expected.

RESPONDENT BURDEN

Both raters took 20 to 25 minutes on average to assess each paper, with a range from 10 to 45 minutes. Not surprisingly, shorter papers took less time than longer ones and the time decreased with increasing familiarity with the topic (stress incontinence surgery) and with the checklist. Both raters felt able to rate the methodological aspects of the studies, though one felt that his lack of knowledge of the topic impaired his performance.

Generally, the raters found the checklist clear with no obvious redundancy of items. Their principal difficulties stemmed from a lack of sufficient definition of some items. For example, "Are the characteristics of the patients included in the study clearly described?" did not make explicit which characteristics ought to have been described. A similar problem arose with the item "Are the interventions of interest described?".

Discussion

This study has shown that it is feasible to develop a checklist that can be used to assess the methodological quality not only of randomised controlled trials but also non-randomised studies. It has also shown that it is possible to produce a checklist that provides a profile of the paper, alerting reviewers to its particular methodological strengths and weaknesses. The performance of the checklist we developed improved considerably after revision of a pilot version. The Quality Index had high internal consistency, good test-retest and inter-rater reliability, and good face and criterion validity. Its performance with randomised controlled trials was as good as another established checklist.[15] There was little difference between its performance with non-randomised and with randomised studies.

The remaining principal area of concern is the assessment of external validity, an aspect that has been ignored in all checklists for randomised controlled trials. There are three possible reasons for the poor reliability of the external validity sub-scale. Firstly, it is made up of only three items (compared with 6–10 for the other sub-scales). Secondly, the construction of the three questions may have been so poor that their meaning was unclear. And thirdly, the raters may have been particularly poor at interpreting the questions correctly. While further revision of the wording of these items needs to be considered, the last explanation seems the most probable reason for the poor results. Despite all four raters having studied epidemiology to Masters level, their recently completed course concentrated on aetiological applications of methods rather than health care evaluation, and thus paid little or no attention to the issue of external validity. This is a methodological aspect that epidemiologists have traditionally ignored in the belief that it is of little importance. While this may be true for aetiological research, there is evidence that the issue of external validity is of importance in health care evaluation.[20–23] It may be an important issue for clinicians trying to interpret the applicability of the results of published studies as they want to know if the procedures, hospital characteristics, and patient sample is relevant to their practice. Therefore inclusion of information relating to external validity may have implications for change in clinical practice. Further development of a checklist should include not only a review of the number and wording of the items making up the external validity sub-scale but also the training of reviewers in identifying the relevant information in studies being assessed.

Further development of the checklist should not be confined to the external validity sub-scale. The reliability of some other items

KEY POINTS

- Evaluation of many areas of health care require the use of non-randomised methods.
- While validated checklists for assessing the quality of randomised controlled trials exist, none exist for non-randomised studies.
- Existing checklists for randomised controlled trials lack sub-scales and ignore the external validity (generalisability) of trials.
- It is feasible to develop a checklist for assessing both randomised and non-randomised studies.
- Further development and testing of the checklist described is required.

(9,14,17,19,26) was poor and these warrant further attention. While greater internal consistency of the sub-scales can be obtained by increasing the number of items, another objective is to minimise respondent burden. A shorter scale could be achieved by the use of factor analysis. In addition, the value of a single Global Score needs to be tested by reviewers making such an assessment before rather than after using the 27 item checklist.

Another methodological issue that requires further investigation is that of weighting. On the basis of current knowledge, we suggest assigning equal weighting to each of the five dimensions. This is based more on the lack of evidence to prioritise one dimension over another rather than on any evidence to suggest each dimension was of the same importance. There are several ways forward. The simplest would be a sensitivity analysis in which the effect of adopting different weightings on the rating and ranking of studies could be observed. A more theoretical approach would entail some form of consensus development among experienced health care epidemiologists in which their views of the relative importance of the five dimensions were considered. Ultimately we need greater knowledge and understanding of the actual impact each dimension has on the effect size of the intervention being studied. Little work of this type has been done so far, though some has recently been reported[24] and more is underway.

This feasibility study made use of only two raters when testing versions of the checklist and the retest for intra-rater reliability was conducted only two weeks after the initial test. Clearly, more rigorous testing with larger numbers of reviewers and a longer period before retesting will be required before the routine use of a version can be encouraged. In addition, it would be of interest to see whether or not familiarity with the clinical aspects of the studies being assessed made any difference to the performance of the checklist. Similarly, the method needs to be applied to other subjects as well as using raters with different levels of epidemiological skill and experience.

Meanwhile, we believe that we have shown that it is feasible to assess the quality of non-randomised studies and that such assessments can be made with the same checklist as is used for randomised studies. While further methodological work is done to improve the checklist in the ways outlined above, we would encourage people to use the existing version rather than either ignoring non-randomised studies or using them but ignoring their methodological quality.

We thank the four raters: Elizabeth Breeze, Megan Landon, Giovanni Leonardi, and Naomi Eaton; and Colin Sanderson, Simon Thompson, and David Leon for their comments on the face and content validity of the pilot version; Donna Lamping for advice on psychometric methods and analysis; and the reviewers for their helpful comments.

Funding: no specific funding.
Conflicts of interest: none.

1 Rosenberg W, Donald A. Evidence based medicine: an approach to clinical problem-solving. *BMJ* 1995;**310**:1122–6.
2 Downs SH, Black NA, Devlin HB, *et al*. Systematic review of the effectiveness and safety of laparoscopic cholecystectomy. *Ann R Coll Surg Engl* 1996;**78**:241–323.
3 Black NA, Downs SH. The effectiveness of surgery for stress incontinence in women: a systematic review. *Br J Urol* 1996;**78**:497–510.
4 Moher D, Jadad AR, Nichol G, *et al.* Assessing the quality of randomised controlled trials: an annotated bibliography of checklists. *Control Clin Trials* 1995;**16**:62–73.
5 Chalmers I. Evaluating the effects of care during pregnancy and childbirth. In: Chalmers I, Enkin M, Keirse MJNC, eds. *Effective care in pregnancy and childbirth.* Oxford: Oxford University Press, 1989.
6 Peto R. Clinical trial reporting. *Lancet* 1996;**348**:894–5.
7 Gardner MJ, Machin D, Campbell MJ. Use of checklists in assessing the statistical content of medical studies. *BMJ* 1986;**292**:810–2.
8 Colditz GA, Miller JN, Mosteller F. How study design affects outcomes in comparisons of therapy I: medical. *Stat Med* 1989;**8**:441–54.
9 Sackett DL. Bias in analytical research. *J Chron Dis* 1979;**32**:51–63.
10 Lionel NDW, Herxheimer A. Assessing reports of therapeutic trials. *BMJ* 1970;**3**:637–40.
11 Badgley RF. An assessment of research methods reported in 103 scientific articles from two Canadian Medical Journals. *Can Med Assoc J* 1961;**85**:246–51.
12 Thomson ME, Kramer MS. Methodological standards for controlled clinical trials of early contact and maternal-infant behaviour. *Pediatrics* 1984;**73**:294–300.
13 Sacks HS, Berrier J, Reitman D, *et al*. Meta-analysis of randomised controlled trials. *N Engl J Med* 1987;**316**:450–5.
14 DerSimonian R, Charette J, McPeek B, *et al*. Reporting on methods in clinical trials. *N Engl J Med* 1982;**306**:1332–7.
15 Standards of Reporting Trials Group. A proposal for structured reporting of randomised controlled trails. *JAMA* 1994;**272**:1926–31.
16 Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;**33**:159–74.
17 Medical Outcomes Trust Scientific Advisory Committee. Instrument review criteria. *Medical Outcomes Trust Bulletin* 1995;**3**:i-iv.
18 Kuder GF, Richardson MW. The theory of the estimation of test reliability. *Psychometrika* 1937;**2**:151–60.
19 Streiner DL, Norman GR. *Health measurement scales. A practical guide to their development and use*. Oxford: Oxford University Press, 1989: 86.
20 Davis CE. Generalising from clinical trials. *Control Clin Trials* 1994;**15**:11–14.
21 Marubini E, Mariani L, Salvadori B, *et al*. Results of a breast-cancer-surgery trial compared with observational data from routine practice. *Lancet* 1996;**347**:1000–3.
22 Bailey KR. Generalising the results of randomized clinical trials. *Control Clin Trials* 1994;**15**:15–23.
23 Harth SC, Thong YH. Sociodemographic and motivational characteristics of parents who volunteer their children for clinical research: a controlled study. *BMJ* 1990;**300**:1372–5.
24 Schulz KF, Chalmers I, Hayes RJ, *et al*. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995;**273**:408–12.

Appendix

Checklist for measuring study quality

Reporting

1. *Is the hypothesis/aim/objective of the study clearly described?*

yes	1
no	0

2. *Are the main outcomes to be measured clearly described in the Introduction or Methods section?*
   If the main outcomes are first mentioned in the Results section, the question should be answered no.

yes	1
no	0

3. *Are the characteristics of the patients included in the study clearly described ?*
   In cohort studies and trials, inclusion and/or exclusion criteria should be given. In case-control studies, a case-definition and the source for controls should be given.

yes	1
no	0

4. *Are the interventions of interest clearly described?*
   Treatments and placebo (where relevant) that are to be compared should be clearly described.

yes	1
no	0

5. *Are the distributions of principal confounders in each group of subjects to be compared clearly described?*
   A list of principal confounders is provided.

yes	2
partially	1
no	0

6. *Are the main findings of the study clearly described?*
   Simple outcome data (including denominators and numerators) should be reported for all major findings so that the reader can check the major analyses and conclusions. (This question does not cover statistical tests which are considered below).

yes	1
no	0

7. *Does the study provide estimates of the random variability in the data for the main outcomes?*
   In non normally distributed data the inter-quartile range of results should be reported. In normally distributed data the standard error, standard deviation or confidence intervals should be reported. If the distribution of the data is not described, it must be assumed that the estimates used were appropriate and the question should be answered yes.

yes	1
no	0

8. *Have all important adverse events that may be a consequence of the intervention been reported?*
   This should be answered yes if the study demonstrates that there was a comprehensive attempt to measure adverse events. (A list of possible adverse events is provided).

yes	1
no	0

9. *Have the characteristics of patients lost to follow-up been described?*
   This should be answered yes where there were no losses to follow-up or where losses to follow-up were so small that findings would be unaffected by their inclusion. This should be answered no where a study does not report the number of patients lost to follow-up.

yes	1
no	0

10. *Have actual probability values been reported(e.g. 0.035 rather than <0.05) for the main outcomes except where the probability value is less than 0.001?*

yes	1
no	0

External validity

All the following criteria attempt to address the representativeness of the findings of the study and whether they may be generalised to the population from which the study subjects were derived.

11. *Were the subjects asked to participate in the study representative of the entire population from which they were recruited?*
    The study must identify the source population for patients and describe how the patients were selected. Patients would be representative if they comprised the entire source population, an unselected sample of consecutive patients, or a random sample. Random sampling is only feasible where a list of all members of the relevant

population exists. Where a study does not report the proportion of the source population from which the patients are derived, the question should be answered as unable to determine.

yes	1
no	0
unable to determine	0

12. *Were those subjects who were prepared to participate representative of the entire population from which they were recruited?*
The proportion of those asked who agreed should be stated. Validation that the sample was representative would include demonstrating that the distribution of the main confounding factors was the same in the study sample and the source population.

yes	1
no	0
unable to determine	0

13. *Were the staff, places, and facilities where the patients were treated, representative of the treatment the majority of patients receive?*
For the question to be answered yes the study should demonstrate that the intervention was representative of that in use in the source population. The question should be answered no if, for example, the intervention was undertaken in a specialist centre unrepresentative of the hospitals most of the source population would attend.

yes	1
no	0
unable to determine	0

Internal validity - bias

14. *Was an attempt made to blind study subjects to the intervention they have received ?*
For studies where the patients would have no way of knowing which intervention they received, this should be answered yes.

yes	1
no	0
unable to determine	0

15. *Was an attempt made to blind those measuring the main outcomes of the intervention?*

yes	1
no	0
unable to determine	0

16. *If any of the results of the study were based on "data dredging", was this made clear?*
Any analyses that had not been planned at the outset of the study should be clearly indicated. If no retrospective unplanned subgroup analyses were reported, then answer yes.

yes	1
no	0
unable to determine	0

17. *In trials and cohort studies, do the analyses adjust for different lengths of follow-up of patients, or in case-control studies, is the time period between the intervention and outcome the same for cases and controls ?*
Where follow-up was the same for all study patients the answer should yes. If different lengths of follow-up were adjusted for by, for example, survival analysis the answer should be yes. Studies where differences in follow-up are ignored should be answered no.

yes	1
no	0
unable to determine	0

18. *Were the statistical tests used to assess the main outcomes appropriate?*
The statistical techniques used must be appropriate to the data. For example non-parametric methods should be used for small sample sizes. Where little statistical analysis has been undertaken but where there is no evidence of bias, the question should be answered yes. If the distribution of the data (normal or not) is not described it must be assumed that the estimates used were appropriate and the question should be answered yes.

yes	1
no	0
unable to determine	0

19. *Was compliance with the intervention/s reliable?*
Where there was non compliance with the allocated treatment or where there was contamination of one group, the question should be answered no. For studies where the effect of any misclassification was likely to bias any association to the null, the question should be answered yes.

yes	1
no	0
unable to determine	0

20. *Were the main outcome measures used accurate (valid and reliable)?*

For studies where the outcome measures are clearly described, the question should be answered yes. For studies which refer to other work or that demonstrates the outcome measures are accurate, the question should be answered as yes.

yes	1
no	0
unable to determine	0

Internal validity - confounding (selection bias)

21. *Were the patients in different intervention groups (trials and cohort studies) or were the cases and controls (case-control studies) recruited from the same population?*
For example, patients for all comparison groups should be selected from the same hospital. The question should be answered unable to determine for cohort and case-control studies where there is no information concerning the source of patients included in the study.

yes	1
no	0
unable to determine	0

22. *Were study subjects in different intervention groups (trials and cohort studies) or were the cases and controls (case-control studies) recruited over the same period of time?*
For a study which does not specify the time period over which patients were recruited, the question should be answered as unable to determine.

yes	1
no	0
unable to determine	0

23. *Were study subjects randomised to intervention groups?*
Studies which state that subjects wererandomised should be answered yes except where method of randomisation would not ensure random allocation. For example alternate allocation would score no because it is predictable.

yes	1
no	0
unable to determine	0

24. *Was the randomised intervention assignment concealed from both patients and health care staff until recruitment was complete and irrevocable?*
All non-randomised studies should be answered no. If assignment was concealed from patients but not from staff, it should be answered no.

yes	1
no	0
unable to determine	0

25. *Was there adequate adjustment for confounding in the analyses from which the main findings were drawn?*
This question should be answered no for trials if: the main conclusions of the study were based on analyses of treatment rather than intention to treat; the distribution of known confounders in the different treatment groups was not described; or the distribution of known confounders differed between the treatment groups but was not taken into account in the analyses. In non-randomised studies if the effect of the main confounders was not investigated or confounding was demonstrated but no adjustment was made in the final analyses the question should be answered as no.

yes	1
no	0
unable to determine	0

26. *Were losses of patients to follow-up taken into account?*
If the numbers of patients lost to follow-up are not reported, the question should be answered as unable to determine. If the proportion lost to follow-up was too small to affect the main findings, the question should be answered yes.

yes	1
no	0
unable to determine	0

Power

27. *Did the study have sufficient power to detect a clinically important effect where the probability value for a difference being due to chance is less than 5%?*
Sample sizes have been calculated to detect a difference of x% and y%.

	Size of *smallest* intervention group	
A	$<n_1$	0
B	n_1–n_2	1
C	n_3–n_4	2
D	n_5–n_6	3
E	n_7–n_8	4
F	n_8+	5