APPENDIX

APPENDIX A SUMMARY TABLE OF THE GENERAL CHARACTERISTICS OF INCULDED STUDIES.

						N		Male				Age			
No.	Author	Year	Study design	Origin of publication	Sub-ethnicity	Case	Control	Case		Control		Case		Control	
								N	%	N	%	Mean	Range	Mean	Range

APPENDIX B SUMMARY TABLE OF THE INFORMATION RELATED TO OUTCOME OF THE INCLUDED STUDIES.

No.	Author	Year	Diagnostic method	Criteria/Definition of case and control				Method of HLA genotyping	
				Case		Control		Case	Control
				Criteria/ Definition	Detail	Criteria/ Definition	Detail		

No.	Author	Year	HLA genotype………………….. Positive/Negative				Quality assessment using Newcastle-Ottawa scale		Quality assessment using risk of bias scale use	
			Case		Control		Scale	Noted	Scale	Noted
			Positive	Negative	Positive	Negative				

APPENDIX C TOOLS FOR ASSESS QUALITY OF THE STUDIES

NEWCASTLE - OTTAWA QUALITY ASSESSMENT SCALE
CASE CONTROL STUDIES

Note: A study can be awarded a maximum of one star for each numbered item within the Selection and Exposure categories. A maximum of two stars can be given for Comparability.

Selection

1) Is the case definition adequate?

a) yes, with independent validation *

b) yes, eg record linkage or based on self-reports

c) no description

2) Representativeness of the cases

a) consecutive or obviously representative series of cases *

b) potential for selection biases or not stated

3) Selection of Controls

a) community controls *

b) hospital controls

c) no description

4) Definition of Controls

a) no history of disease (endpoint) *

b) no description of source

Comparability

1) Comparability of cases and controls on the basis of the design or analysis

a) study controls for _______________ (Select the most important factor.) *

b) study controls for any additional factor * (This criteria could be modified to indicate specific control for a second important factor.)

Exposure

1) Ascertainment of exposure

a) secure record (eg surgical records) *

b) structured interview where blind to case/control status *

c) interview not blinded to case/control status

d) written self-report or medical record only

e) no description

2) Same method of ascertainment for cases and controls

a) yes *

b) no

3) Non-Response rate

a) same rate for both groups *

b) non respondents described

c) rate different and no designation

NEWCASTLE - OTTAWA QUALITY ASSESSMENT SCALE

COHORT STUDIES

Note: A study can be awarded a maximum of one star for each numbered item within the Selection and Outcome categories. A maximum of two stars can be given for Comparability

Selection

1) Representativeness of the exposed cohort

a) truly representative of the average _______________ (describe) in the community *

b) somewhat representative of the average ______________ in the community *

c) selected group of users e.g. nurses, volunteers

d) no description of the derivation of the cohort

2) Selection of the non-exposed cohort

a) drawn from the same community as the exposed cohort *

b) drawn from a different source

c) no description of the derivation of the non-exposed cohort

3) Ascertainment of exposure

a) secure record (e.g. surgical records) *

b) structured interview *

c) written self-report

d) no description

4) Demonstration that outcome of interest was not present at start of study

a) yes *

b) no

Comparability

1) Comparability of cohorts on the basis of the design or analysis

a) study controls for _____________ (select the most important factor) *

b) study controls for any additional factor * (This criteria could be modified to indicate specific control for a second important factor.)

Outcome

1) Assessment of outcome

   a) independent blind assessment *

   b) record linkage *

   c) self-report

   d) no description

2) Was follow-up long enough for outcomes to occur

   a) yes (select an adequate follow up period for outcome of interest) *

   b) no

3) Adequacy of follow up of cohorts

   a) complete follow up - all subjects accounted for *

   b) subjects lost to follow up unlikely to introduce bias - small number lost - > ____ % (select an adequate %) follow up, or description provided of those lost) *

   c) follow up rate < __% (select an adequate %) and no description of those lost

   d) no statement

APPENDIX D TOOLS FOR ASSESS QUALITY OF THE STUDIES

The Cochrane Collaboration's tool for assessing risk of bias

Domain	**Description**	**Review authors' judgement**
Sequence generation	Describe the method used to generate the allocation sequence in sufficient detail to allow an assessment of whether it should produce comparable groups.	Was the allocation sequence adequately generated?
Allocation concealment	Describe the method used to conceal the allocation sequence in sufficient detail to determine whether intervention allocations could have been foreseen in advance of, or during, enrolment.	Was allocation adequately concealed?
Blinding of participants, personnel and outcome assessors *Assessments should be made for each main outcome (or class of outcomes)*	Describe all measures used, if any, to blind study participants and personnel from knowledge of which intervention a participant received. Provide any information relating to whether the intended blinding was effective.	Was knowledge of the allocated intervention adequately prevented during the study?
Incomplete outcome data *Assessments should be made for each main outcome (or class of outcomes)*	Describe the completeness of outcome data for each main outcome, including attrition and exclusions from the analysis. State whether attrition and exclusions were reported, the numbers in each intervention group (compared with total randomized participants), reasons for attrition/exclusions where reported, and any re-inclusions in analyses performed by the review authors.	Were incomplete outcome data adequately addressed?
Selective outcome reporting	State how the possibility of selective outcome reporting was examined by the review authors, and what was found.	Are reports of the study free of suggestion of selective outcome reporting?
Other sources of bias	State any important concerns about bias not addressed in the other domains in the tool. If particular questions/entries were pre-specified in the review's protocol, responses should be provided for each question/entry.	Was the study apparently free of other problems that could put it at a high risk of bias?

Possible approach for *summary assessments* outcome (across domains) within and across studies

Risk of bias	**Interpretation**	**Within a study**	**Across studies**
Low risk of bias	Plausible bias unlikely to seriously alter the results.	Low risk of bias for all key domains.	Most information is from studies at low risk of bias.
Unclear risk of bias	Plausible bias that raises some doubt about the results	Unclear risk of bias for one or more key domains.	Most information is from studies at low or unclear risk of bias.
High risk of bias	Plausible bias that seriously weakens confidence in the results.	High risk of bias for one or more key domains.	The proportion of information from studies at high risk of bias is sufficient to affect the interpretation of the results.

Criteria for judging risk of bias in the 'Risk of bias' assessment tool

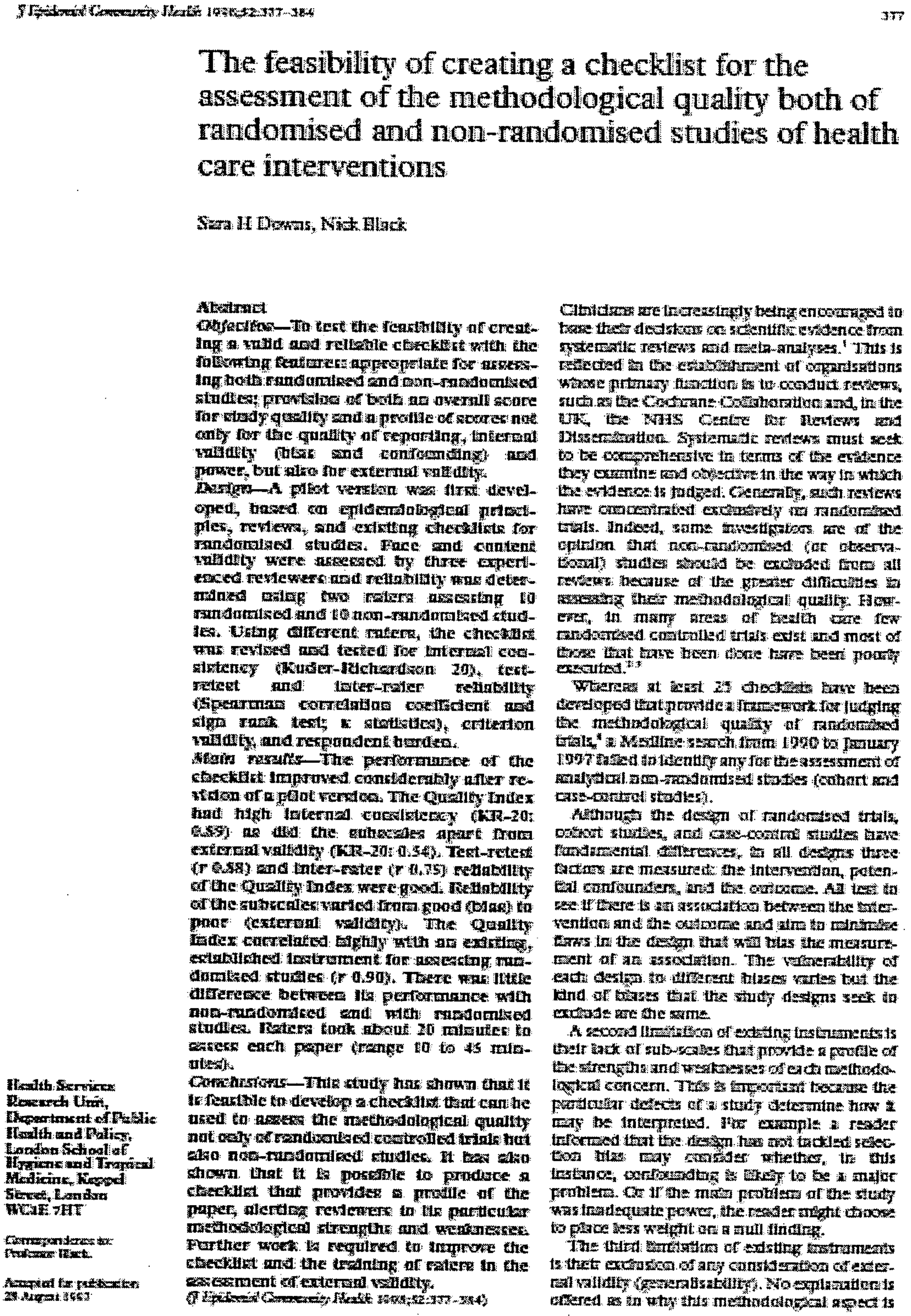
SEQUENCE GENERATION **Was the allocation sequence adequately generated? [Short form: *Adequate sequence generation*?]**	
Criteria for a judgement of 'YES' (i.e. low risk of bias).	The investigators describe a random component in the sequence generation process such as: ▪ Referring to a random number table; Using a computer random number generator; Coin tossing; Shuffling cards or envelopes; Throwing dice; Drawing of lots; Minimization*. *Minimization may be implemented without a random element, and this is considered to be equivalent to being random.
Criteria for the judgement of 'NO' (i.e. high risk of bias).	The investigators describe a non-random component in the sequence generation process. Usually, the description would involve some systematic, non-random approach, for example: ▪ Sequence generated by odd or even date of birth; ▪ Sequence generated by some rule based on date (or day) of admission; ▪ Sequence generated by some rule based on hospital or clinic record number. Other non-random approaches happen much less frequently than the systematic approaches mentioned above and tend to be obvious. They usually involve judgement or some method of non-random categorization of participants, for example: ▪ Allocation by judgement of the clinician; ▪ Allocation by preference of the participant; ▪ Allocation based on the results of a laboratory test or a series of tests; ▪ Allocation by availability of the intervention.
Criteria for the judgement of 'UNCLEAR' (uncertain risk of bias).	Insufficient information about the sequence generation process to permit judgement of 'Yes' or 'No'.
ALLOCATION CONCEALMENT **Was allocation adequately concealed? [Short form: *Allocation concealment*?]**	
Criteria for a judgement of 'YES' (i.e. low risk of bias).	Participants and investigators enrolling participants could not foresee assignment because one of the following, or an equivalent method, was used to conceal allocation: ▪ Central allocation (including telephone, web-based, and pharmacy-controlled, randomization); ▪ Sequentially numbered drug containers of identical appearance; ▪ Sequentially numbered, opaque, sealed envelopes.
Criteria for the judgement of 'NO' (i.e. high risk of bias).	Participants or investigators enrolling participants could possibly foresee assignments and thus introduce selection bias, such as allocation based on: ▪ Using an open random allocation schedule (e.g. a list of random numbers); ▪ Assignment envelopes were used without appropriate safeguards (e.g. if envelopes were unsealed or non-opaque or not sequentially numbered); ▪ Alternation or rotation; ▪ Date of birth; ▪ Case record number; ▪ Any other explicitly unconcealed procedure.

Criteria for the judgement of 'UNCLEAR' (uncertain risk of bias).	Insufficient information to permit judgement of 'Yes' or 'No'. This is usually the case if the method of concealment is not described or not described in sufficient detail to allow a definite judgement – for example if the use of assignment envelopes is described, but it remains unclear whether envelopes were sequentially numbered, opaque and sealed.
BLINDING OF PARTICIPANTS, PERSONNEL AND OUTCOME ASSESSORS **Was knowledge of the allocated interventions adequately prevented during the study? [Short form: *Blinding*?]**	
Criteria for a judgement of 'YES' (i.e. low risk of bias).	Any one of the following: ▪ No blinding, but the review authors judge that the outcome and the outcome measurement are not likely to be influenced by lack of blinding; ▪ Blinding of participants and key study personnel ensured, and unlikely that the blinding could have been broken; ▪ Either participants or some key study personnel were not blinded, but outcome assessment was blinded and the non-blinding of others unlikely to introduce bias.
Criteria for the judgement of 'NO' (i.e. high risk of bias).	Any one of the following: ▪ No blinding or incomplete blinding, and the outcome or outcome measurement is likely to be influenced by lack of blinding; ▪ Blinding of key study participants and personnel attempted, but likely that the blinding could have been broken; ▪ Either participants or some key study personnel were not blinded, and the non-blinding of others likely to introduce bias.
Criteria for the judgement of 'UNCLEAR' (uncertain risk of bias).	Any one of the following: ▪ Insufficient information to permit judgement of 'Yes' or 'No'; ▪ The study did not address this outcome.
INCOMPLETE OUTCOME DATA **Were incomplete outcome data adequately addressed? [Short form: *Incomplete outcome data addressed*?]**	
Criteria for a judgement of 'YES' (i.e. low risk of bias).	Any one of the following: ▪ No missing outcome data; ▪ Reasons for missing outcome data unlikely to be related to true outcome (for survival data, censoring unlikely to be introducing bias); ▪ Missing outcome data balanced in numbers across intervention groups, with similar reasons for missing data across groups; ▪ For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk not enough to have a clinically relevant impact on the intervention effect estimate; ▪ For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes not enough to have a clinically relevant impact on observed effect size; ▪ Missing data have been imputed using appropriate methods.
Criteria for the judgement of 'NO' (i.e. high risk of bias).	Any one of the following: ▪ Reason for missing outcome data likely to be related to true outcome, with either imbalance in numbers or reasons for missing data across intervention groups; ▪ For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk enough to induce clinically relevant bias in intervention effect estimate; ▪ For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes enough to induce clinically relevant bias in observed effect size; ▪ 'As-treated' analysis done with substantial departure of the intervention received from that assigned at randomization; ▪ Potentially inappropriate application of simple imputation.

Criteria for the judgement of 'UNCLEAR' (uncertain risk of bias).	Any one of the following: ▪ Insufficient reporting of attrition/exclusions to permit judgement of 'Yes' or 'No' (e.g. number randomized not stated, no reasons for missing data provided); ▪ The study did not address this outcome.
SELECTIVE OUTCOME REPORTING **Are reports of the study free of suggestion of selective outcome reporting? [Short form: *Free of selective reporting*?]**	
Criteria for a judgement of 'YES' (i.e. low risk of bias).	Any of the following: ▪ The study protocol is available and all of the study's pre-specified (primary and secondary) outcomes that are of interest in the review have been reported in the pre-specified way; ▪ The study protocol is not available but it is clear that the published reports include all expected outcomes, including those that were pre-specified (convincing text of this nature may be uncommon).
Criteria for the judgement of 'NO' (i.e. high risk of bias).	Any one of the following: ▪ Not all of the study's pre-specified primary outcomes have been reported; ▪ One or more primary outcomes is reported using measurements, analysis methods or subsets of the data (e.g. subscales) that were not pre-specified; ▪ One or more reported primary outcomes were not pre-specified (unless clear justification for their reporting is provided, such as an unexpected adverse effect); ▪ One or more outcomes of interest in the review are reported incompletely so that they cannot be entered in a meta-analysis; ▪ The study report fails to include results for a key outcome that would be expected to have been reported for such a study.
Criteria for the judgement of 'UNCLEAR' (uncertain risk of bias).	Insufficient information to permit judgement of 'Yes' or 'No'. It is likely that the majority of studies will fall into this category.
OTHER POTENTIAL THREATS TO VALIDITY **Was the study apparently free of other problems that could put it at a risk of bias? [Short form: *Free of other bias*?]**	
Criteria for a judgement of 'YES' (i.e. low risk of bias).	The study appears to be free of other sources of bias.
Criteria for the judgement of 'NO' (i.e. high risk of bias).	There is at least one important risk of bias. For example, the study: ▪ Had a potential source of bias related to the specific study design used; or ▪ Stopped early due to some data-dependent process (including a formal-stopping rule); or ▪ Had extreme baseline imbalance; or ▪ Has been claimed to have been fraudulent; or ▪ Had some other problem.
Criteria for the judgement of 'UNCLEAR' (uncertain risk of bias).	There may be a risk of bias, but there is either: ▪ Insufficient information to assess whether an important risk of bias exists; or ▪ Insufficient rationale or evidence that an identified problem will introduce bias.

APPENDIX E TOOLS FOR ASSESS QUALITY OF THE STUDIES

(http://jech.bmj.com/content/52/6/377.full.pdf+html)



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

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

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)

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 8] and existing checklists for the assessment of randomised controlled trials.[4 10–14] 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 (25) 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.[9] 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.[18] 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	4.2	3-5
Confounding	3.6	1-6	1.3	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	0.82	0.81
Confounding (6)	0.69	0.74	0.48
Bias (7)	0.78	0.85	0.75
External validity (3)	0.54	0.68	0.15
Internal + external validity (16)	0.72	0.80	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	0.75	0.84
Report	0.84	0.65	0.88	0.54	0.73	0.92
Confounding	0.69	0.0008	0.77	0.10	0.53	0.31
Bias	0.80	0.15	0.85	0.11	0.86	0.89
External validity	0.37	0.33	0.23	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	0.50	25
	2	0.39	35	0.28	35
	3	0.80	10	0.48	25
	4	0.85	5	0.82	30
	5	0.64	10	0.38	30
	6	0.53	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	0.69	15
	17	0.06	40	0.30	15
	18	0.35	30	0.22	20
	19	−0.05	10	0.00	5
	20	0.76	10	0.55	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	0.65	20
	25	0.78	10	0.47	20
	26	0.27	25	0.23	15

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 ($\kappa = -0.18$) and only one of the six items comprising the Confounding sub-scale (item 26; $\kappa = 0.27$). 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 SHTG (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	0.79	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	0.73	0.72	0.90	0.33
External validity (3)	−0.14	0.71	−0.25	0.92	0.92	0.61
SHTG (32)	NA	NA	0.82	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 SHTG's instrument was similar to that obtained for the new Quality Index. The proportion of SHTG'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 SHTG (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.[illegible] 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 Brown, Megan Landon, Gerrard Lavard, and Naomi Fulop; 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 randomized 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 Prie R. Clinical trial reporting. *Lancet* 1994;344:894–5.
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 randomized 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 structural reporting of randomized controlled trials. *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;3:1–4.
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: 65.
20 Davis CE. Generalizing 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. Generalizing 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 were randomised 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

APPENDIX F THE DERSIMONIAN AND NAN LAIRD METHOD

Meta-Analysis in Clinical Trials*

Rebecca DerSimonian and Nan Laird

ABSTRACT: This paper examines eight published reviews each reporting results from several related trials. Each review pools the results from the relevant trials in order to evaluate the efficacy of a certain treatment for a specified medical condition. These reviews lack consistent assessment of homogeneity of treatment effect before pooling. We discuss a random effects approach to combining evidence from a series of experiments comparing two treatments. This approach incorporates the heterogeneity of effects in the analysis of the overall treatment efficacy. The model can be extended to include relevant covariates which would reduce the heterogeneity and allow for more specific therapeutic recommendations. We suggest a simple noniterative procedure for characterizing the distribution of treatment effects in a series of studies.

KEY WORDS: *random effects model, heterogeneity of treatment effects, distribution of treatment effects, covariate information*

INTRODUCTION

Meta-analysis is defined here as the statistical analysis of a collection of analytic results for the purpose of integrating the findings. Such analyses are becoming increasingly popular in medical research where information on efficacy of a treatment is available from a number of clinical studies with similar treatment protocols. If considered separately, any one study may be either too small or too limited in scope to come to unequivocable or generalizable conclusions about the effect of treatment. Combining the findings across such studies represents an attractive alternative to strengthen the evidence about the treatment efficacy.

The main difficulty in integrating the results from various studies stems from the sometimes diverse nature of the studies, both in terms of design and methods employed. Some are carefully controlled randomized experi-

Yale University, School of Medicine, New Haven, Connecticut (R.D.); Harvard University, School of Public Health, Boston, Massachusetts (N.L.)

*This research was supported by grant CA09424-03 from the National Cancer Institute and grant GM-29745 from the National Institute of Health. We are grateful to Frederick Mosteller, Tom Louis, and Katherine Halvorsen for critical readings of various drafts, encouragement, and advice.

Address reprint request to: Rebecca DerSimonian, Yale University School of Medicine, P.O. Box 3333, New Haven, CT 06510.

Received March 25, 1986; accepted April 7, 1986.

52 Vanderbilt Ave., New York, New York 10017

0197-2456/86/$3.50

ments while others are less well controlled. Because of differing sample sizes and patient populations, each study has a different level of sampling error as well. Thus one problem in combining studies for integrative purposes is the assignment of weights that reflect the relative "value" of the information provided in a study. A more difficult issue in combining evidence is that one may be using incommensurable studies to answer the same question. Armitage [1] emphasizes the need for careful consideration of methods in drawing inferences from heterogeneous but logically related studies. In this setting, the use of a regression analysis to characterize differences in study outcomes may be more appropriate [2].

This paper discusses an approach to meta-analysis which addresses these two problems. In this approach, we assume that there is a distribution of treatment effects and utilize the observed effects from individual studies to estimate this distribution. The approach allows for treatment effects to vary across studies and provides an objective method for weighting that can be made progressively more general by incorporating study characteristics into the analysis. We illustrate the use of this model in several examples, and based on the empirical evidence, suggest a simple noniterative procedure for testing and estimation.

DATABASE

In a systematic search of the first ten issues published in 1982 of each of four weekly journals (*NEJM*, *JAMA*, *BMJ*, and *Lancet*), Halvorsen [3] found only one article (out of 589) that considered combining results using formal statistical methods. Our data consist of an ad hoc collection of such articles from the medical literature found through references provided by colleagues and through bibliographic references in articles already located [4–11]. The method we propose applies to several additional articles that have come to our attention since our original analyses [12–14].

We examine in detail eight review articles each reporting results from several related trials. Each review pools the results from the relevant trials in order to evaluate the efficacy of a certain treatment for a specified medical condition. In most of these reviews the original investigators pool the results from the relevant trials and estimate an overall treatment effect without first checking whether the treatment effect across the trials is constant. Others exclude some trials and combine the results only from trials that are similar in design and implementation. The investigators who do check for homogeneity of treatment effect before pooling use different criteria to assess this homogeneity. With two exceptions [6,11], the reviews consider randomized trials only. The two reviews that include nonrandomized studies analyze the data from the two groups of studies (randomized and nonrandomized) separately. In this study, we restrict our attention to the results of randomized trials only. We first describe the eight reviews identifying each by its first author, and in Table 1 summarize the methods used in each review:

Winship: A review of eight trials that compare the healing rates in duodenal ulcer patients treated with cimetidine or placebo therapy [4].

Table 1 The Methods and Outcome Measures Used in the Original Reviews

	Outcome measure	Overall effect estimate	Test of homogeneity
Winship	difference in proportions	pooled raw data	—
Conn	difference in proportions	pooled raw data	—
Miao	difference in proportions	weighted average	Chi-Square Test (Q_w)
DeSilva	relative risk	Mantel–Haenszel estimate for pooled relative risk	—
Stjernsward	difference in proportions	minimum and maximum	—
Baum	difference in proportions	pooled raw data	Gilbert, McPeek, and Mosteller estimate of variance [15]
Peto	difference in proportions	—	Mentions lack of heterogeneity
Chalmers	difference in proportions	unweighted average	—

Conn: A review of nine trials that compare the survival rates in alcoholic hepatitis patients with steroids or control therapy [5].

Miao: A review of six trials that compare gastric with sham freezing in the treatment of duodenal ulcer [6]. In addition, this review considers 14 observational and two controlled but nonrandomized studies.

DeSilva: A review of six trials that evaluate the effect of lignocaine on the incidence of ventricular fibrillation in acute myocardial infarction [7]. This study originally considered 15 trials but because the trials vary widely in treatment schedules and doses, some criteria for adequacy of treatment are established and only six trials that fulfill these requirements are analyzed.

Stjernsward: A review of five trials that compare the 5-year survival rates of patients with cancer of the breast treated with surgery plus radiotherapy or surgery alone [8].

Baum: A review of 26 trials that evaluate the efficacy of antibiotics in the prevention of wound infection following colon surgery [9].

Peto: A review of six trials that evaluate the efficacy of aspirin in the prevention of secondary mortality in persons recovered from myocardial infarction [10].

Chalmers: A review of a number of trials that evaluate the efficacy of anticoagulants in the treatment of acute myocardial infarction [11]. Data from 18 surveys employing historical controls (HCT), eight studies employing alternately assigned controls (ACT), and six randomized controlled trials (RCT) are given. Three endpoints, total case fatality rates, case fatality rates excluding early deaths, and thromboembolism rates are considered, although not all studies report all three endpoints. Here we consider thromboembolism and total case fatality rates in the RCTs only. The results from randomized trials are compared to those of nonrandomized ones (HCTs and ACTs) in Laird and DerSimonian [16].

METHODS

We consider the problem of combining information from a series of k comparative clinical trials, where the data from each trial consist of the number of patients in treatment and control groups, n_T and n_C, and the proportion of patients with some event in each of the groups, r_T and r_C. Letting i index the trials, we assume that the numbers of patients with the event in each of the study groups are independent binomial random variables with associated probabilities p_{Ti} and p_{Ci}, $i = 1, \ldots, k$. The basic idea of the random effects approach is to parcel out some measure of the observed treatment effect in each study, say y_i, into two additive components: the true treatment effect, θ_i, and the sampling error, e_i. The variance of e_i is the sample variance, s_i^2, and is usually calculated from the data of the ith observed sample. The true treatment effect associated with each trial will be influenced by several factors, including patient characteristics as well as design and execution of the study. To explicitly account for the variation in the true effects, the model assumes

$$\theta_i = \mu + \delta_i$$

where θ_i is the true treatment effect in the ith study, μ is the mean effect for a population of possible treatment evaluations, and δ_i is the deviation of the ith study's effect from the population mean. We regard the trials considered as a sample from this population and use the observed effects to estimate μ as well as the population variance [$\mathrm{var}(\delta) = \Delta^2$]. Here, Δ^2 represents both the degree to which treatment effects vary across experiments as well as the degree to which individual studies give biased assessments of treatment effects.

The model just described can thus be characterized by two distinct sampling stages. First we sample a study from a population of possible studies with mean treatment effect μ and variance in treatment effects of Δ^2. Then we sample observations in the ith study with underlying treatment effect θ_i.

One issue which deserves some attention is the specification of treatment effect, θ_i. Three commonly used measures are the risk difference, $p_{Ti} - p_{Ci}$, the relative risk, p_{Ti}/p_{Ci}, and the relative odds, $[p_{Ti}/(1 - p_{Ti})]/[p_{Ci}/(1 - p_{Ci})]$. The relative odds is popular because of its suitability in retrospective or case control studies, and because it has some interesting mathematical properties. In particular, if we assume a constant relative odds ($\theta_i = \mu$ or $\Delta^2 = 0$), then the Mantel–Haenszel statistic is optimal for testing H_0: $\mu = 1$, and there is considerable literature on efficient estimates of μ and on methods for testing H_0: $\theta_1 = \theta_2 = \ldots = \theta_k$. Despite these advantages, the relative odds (and the closely related relative risk) suffers in interpretability. By far the most intuitively appealing measure for trials of clinical efficacy is the risk difference, since it measures actual gains which can be expected in terms of percentages of patients treated. Besides relevance of the measure and statistical efficiency, it is also desirable to choose a measure which is nearly constant over studies, so that the effect of heterogeneity is minimized. Unless there is no treatment effect at all ($p_{Ti} = p_{Ci}$ for all i), constancy of treatment effect in one scale (say $p_{Ti}/p_{Ci} = A$ for all i) implies variation across studies in another ($p_{Ti} - p_{Ci}$, say). Thus it is conceivable that the wrong choice of scale could imply heterogeneity in treatment effects which would not exist if a different measure were used. However, this is not likely to happen in practice unless there is a very wide range in the control rates (p_{Ci}) or all the rates are very close to zero (or one). In such cases, one might want to do the analysis in both the relative odds and risk difference scales.

HOMOGENEITY OF TREATMENT EFFECT

To evaluate constancy of treatment effect across strata, we use a large sample test based on the statistic $Q = \sum_i w_i(y_i - \bar{y}_w)^2$, where y_i is the ith treatment effect estimate, $\bar{y}_w = \sum_i w_i\, y_i/\sum_i w_i$ is the weighted estimator of treatment effect, and w_i is the inverse of the ith sampling variance. The test statistic Q is the sum of squares of the treatment effect about the mean where the ith square is weighted by the reciprocal of the estimated variance. Under the null hypothesis, Q is approximately a χ^2 statistic with $k - 1$ degrees of freedom; thus, when each study has a large sample size relative to the number of strata, Q may be used to test H_0: $\Delta^2 = 0$.

$$\Delta_w^2 = \max \{0, \{Q_w - (k - 1)\} / [\sum_i w_i - (\sum_i w_i^2 / \sum_i w_i)]\},$$

where Q_w, $\bar{y}_w$, w_i are as described above. The weighted least squares or Cochran's [19] semiweighted estimator of μ is

$$\mu_w = \sum_i w_i^* y_i / \sum_i w_i^*, \tag{2}$$

where

$$w_i^* = (w_i^{-1} + \Delta_w^2)^{-1}. \tag{3}$$

The asymptotic standard error of μ_w is

$$\text{s.e.}\ (\mu_w) = (\sum_i w_i^*)^{-1/2}. \tag{4}$$

When the sampling variances are assumed to be equal, these equations reduce to:

$$\Delta_u^2 = \max [0, \{\sum_i (y_i - \bar{y})^2 / (k - 1)\} - s^2],$$

$$\mu_u = \bar{y},$$

and

$$\text{s.e.}\ (\mu_u) = [(s^2 + \Delta_u^2) / k]^{1/2},$$

where

$$\bar{y} = \sum_i y_i / k \qquad \text{and} \qquad s^2 = \sum_i s_i^2 / k.$$

Rao et al. [20] derive Δ_u^2 from an unweighted sum of squares procedure and show that it is also the Minque estimator when the sampling variances are all equal. The unweighted mean, μ_u, is equivalent to the estimate of the treatment effect in reviews that use the average difference in proportions to assess the overall treatment efficacy.

With an additional assumption that y_i is $N(\theta_i, s_i^2)$ and θ_i is $N(\mu, \Delta^2)$, we also compute maximum likelihood (ML) and restricted maximum likelihood (REML) estimates and compare them to the noniterative ones. The maximum likelihood estimates of the unknown parameters are those values that maximize the probability density function of the data. In REML estimation, the likelihood to be maximized is slightly modified to adjust for μ and Δ^2 being estimated from the same data. The REML estimators are the iterative equivalents of the weighted estimators above. Both ML and REML estimates of μ and its s.e. take the form given in equations (2) and (4) with weights given in (3), but differ in the way Δ^2 is estimated.

The ML estimating equations are given in Rao et al. [20] and the REML equations are reviewed by Harville [21]. For implementing the ML or REML procedures, we use the EM algorithm [22] which is an iterative procedure for computing maximum likelihood estimates appropriate when the observations can be viewed as incomplete data.

RESULTS

Homogeneity of Treatment Effect

We present the statistics for testing homogeneity of treatment effect in Table 2. For these reviews Q_w, the weighted statistic in the difference scale, and Q_L, the analogous statistic in the log odds scale, imply similar conclusions about the constancy of treatment effect. The assumption of homogeneity holds in the reviews by DeSilva [7], Stjernsward [8], Peto [10], and in Chalmers' [11] randomized controlled trials (case fatality rates). In the remaining five sets of trials, the evidence suggests heterogeneity of treatment effect irrespective of the scale of measurement.

The review by Peto [10] mentions lack of heterogeneity in treatment effects across trials. For this review, the Q_u statistic supports the homogeneity assumption (p value = 0.65), whereas both Q_w and Q_L support that assumption only marginally (p value = 0.12). Baum et al. [9] estimate the variability in treatment differences using the method of Gilbert et al. [15] and conclude that relative to within study variation (assumed equal for all studies), between study variation is negligible. This qualitative assessment is not consistent with the results of Table 2 where a common treatment effect across the trials in this review does not seem to hold in either scale of measurement. The third review which includes a test of homogeneity before pooling the results [6] uses a slightly modified version of Q_w to test this hypothesis and the conclusion of lack of homogeneity agrees with the result in Table 2.

As in the review by Peto [10], Q_u and Q_w imply different conclusions about the homogeneity of treatment effect in the review by Winship [4]. In this review also, the method assuming equal weights (Q_u) implies homogeneity of effect while the weighted one implies the opposite.

These results emphasize that the variation in the treatment effect across several trials is often not negligible and should be incorporated into the anal-

Table 2 Test of Homogeneity[a]

	df[c]	Q_w[d]	Q_L[e]	Q_u[f]
Winship	7	15.2[b]	15.6[b]	7.9
Conn	8	15.6[b]	20.6[b]	19.3[b]
Miao	5	21.7[b]	18.8[b]	20.9[b]
DeSilva	5	9.1	4.4	7.3
Stjernsward	4	2.1	2.2	2.7
Baum	25	40.4[b]	35.2[b]	35.3[b]
Peto	5	9.0	9.2	3.5
Chalmers				
Thromboembolism	5	12.3[b]	10.3[b]	10.6[b]
Case fatality rates	5	3.5	2.4	1.8

[a] Figures in Tables 2–4 are based on data available at the time of review publication.

[b] p value <0.10.

[c] Degrees of freedom.

[d] Q statistic in difference scale (unequal weights).

[e] Q statistic in log odds scale (unequal weights).

[f] Q statistic in difference scale (equal weights).

ysis of the overall treatment efficacy. Lack of homogeneity holds both when the treatment effect is the difference in proportions and when it is the log odds. The unweighted statistic which assigns an equal weight to each study may not be appropriate for testing homogeneity when differences in sample sizes and/or underlying proportions across studies are large.

Estimation

For all four methods of estimation we present estimates of μ and its s.e. in Table 3, and estimates of Δ^2 in Table 4. The estimates of Δ^2, and s.e. ($\hat{\mu}$) are quite similar in the weighted noniterative method, maximum likelihood, and restricted maximum likelihood procedures. The Δ^2s from these three methods are zero or nearly so in the reviews by DeSilva [7], Stjernsward [8], Peto [10], and Chalmers' [11] randomized trials (case fatality rates). These same reviews have Q statistics that are small relative to their degrees of freedom (Table 2). The weighted method and the REML estimation procedures consistently yield slightly higher values of Δ^2 than the ML procedure. This is because both these procedures adjust for μ and Δ^2 being estimated from the same data whereas the ML procedure does not. The estimates of μ and its s.e. from these three procedures are expected to be similar since the estimates of Δ^2 are almost equal.

Comparing the unweighted method of moments with the other three methods, we find that the estimates for Δ^2 from this method differ, and sometimes differ widely, from the estimates of the other three methods but without any consistent pattern. The estimates of μ and its s.e. from the unweighted method also differ from the estimates of the other three methods and these differences are not necessarily due to the differences in Δ^2s. In Chalmers' [11] randomized trials (case fatality rates), for instance, even when Δ^2 is zero for all four methods, the estimate of μ is 0.042 (s.e. = 0.024) for the unweighted method while it is 0.029 (s.e. = 0.012) for the other three methods. The original reviewers report the unweighted average of the observed rate differences

Table 3 Estimated Overall Effects and Their Standard Errors[a]

	μ_u^b	μ_w^c	μ_M^d	μ_R^e
Winship	0.406 (0.046)	0.389 (0.058)	0.384 (0.053)	0.387 (0.056)
Conn	0.102 (0.092)	0.075 (0.072)	0.070 (0.063)	0.073 (0.069)
Miao	0.077 (0.125)	0.094 (0.111)	0.095 (0.106)	0.093 (0.118)
DeSilva	0.026 (0.019)	0.027 (0.019)	0.026 (0.017)	0.027 (0.019)
Stjernsward	0.046 (0.020)	0.041 (0.018)	0.041 (0.018)	0.041 (0.018)
Baum	0.203 (0.031)	0.208 (0.026)	0.208 (0.025)	0.208 (0.026)
Peto	0.018 (0.008)	0.015 (0.008)	0.014 (0.008)	0.015 (0.008)
Chalmers				
Thromboembolism	0.102 (0.036)	0.079 (0.020)	0.078 (0.017)	0.078 (0.020)
Case fatality rates	0.042 (0.024)	0.029 (0.012)	0.029 (0.012)	0.029 (0.012)

[a] Figures in parentheses represent the standard errors of the corresponding estimates.

[b] Noniterative estimates with equal weights.

[c] Noniterative estimates with weights to reflect unequal variances.

[d] Maximum likelihood estimates.

[e] Restricted maximum likelihood estimates.

the other hand, the unweighted method which ignores differences in sample sizes yields estimates that often differ from the estimates of the other methods.

A problem in pooling data we have not addressed here is that of publication bias. This problem relates to studies being executed, but not reported, usually because treatment effect has not been found. Reviewers generally recount those studies that appear to be worthwhile and discount those that are unpublished or are not in agreement with a favored group of studies. The method we describe here represents a systematic, quantitative pooling of available data to resolve controversies about a treatment effect. With each individual controversy, unpublished information may be elicited and along with recent findings the method can be used to update the results.

In all our work we assume that the sampling variances are known, although in reality we estimate them from the data. Further research needs to be done in this area as there are alternative estimators that might be preferable to the ones we use. For instance, if the sample sizes in each study are small, then sampling variances based on pooled estimates of the proportions in the treatment and control groups might be better than the ones based on estimates of proportions from the individual studies. Another alternative is to shrink the individual proportions towards a pooled estimate before calculating the variances. Further investigation is needed before one single method emerges as superior.

REFERENCES

1. Armitage P: Controversies and achievements in clinical trials. Controlled Clin Trials 5: 67–72, 1984
2. DerSimonian R, Laird N: Evaluating the effect of coaching on SAT Scores: a meta-analysis. Harvard Ed Rev 53: 1–15, 1983
3. Halvorsen K: Combining results from independent investigations: meta-analysis in medical research. In: Medical Uses of Statistics, Bailar JC, Mosteller F, Eds. Boston: New England Journal of Medicine (in press)
4. Winship D: Cimetidine in the treatment of duodenal ulcer. Gastroenterology 74: 402–406, 1978
5. Conn H: Steroid treatment of alcoholic hepatitis. Gastroenterology 74: 319–326, 1978
6. Miao L: Gastric freezing: an example of the evaluation of medical therapy by randomized clinical trials. In: Costs, Risks, and Benefits of Surgery, Bunker JP, Barnes BA, Mosteller F, Eds. New York: Oxford University Press, 1977, pp. 198–211
7. DeSilva RA, Hennekens CH, Lown B, Casscells W: Lignocaine prophylaxis in acute myocardial infarction: An evaluation of randomized trials. Lancet ii: 855–858, 1981
8. Stjernsward J: Decreased survival related to irradiation post-operatively in early operable breast cancer. Lancet ii: 1285–1286, 1974
9. Baum ML, Anish DS, Chalmers TC, Sacks HS, Smith H, Fagerstrom, RM: A survey of clinical trials of antibiotic prophylaxis in colon surgery: evidence against further use of no-treatment controls. N Engl J Med 305:795–799, 1981
10. Peto R: Aspirin after myocardial infarction. Lancet i: 1172–1173, 1980 (unsigned editorial)
11. Chalmers TC, Matta RJ, Smith H, Kunzler AM: Evidence favoring the use of anticoagulants in the hospital phase of acute myocardial infarction. N Engl J Med 297: 1091–1096, 1977