Published by Oxford University Press on behalf of the International Epidemiological Association

International Journal of Epidemiology 2007;**36**:666–676
doi:10.1093/ije/dym018

Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography

Simon Sanderson,[1]* Iain D Tatt[2,4] and Julian PT Higgins[3]

Accepted 29 January 2007

Background Assessing quality and susceptibility to bias is essential when interpreting primary research and conducting systematic reviews and meta-analyses. Tools for assessing quality in clinical trials are well-described but much less attention has been given to similar tools for observational epidemiological studies.

Methods Tools were identified from a search of three electronic databases, bibliographies and an Internet search using Google®. Two reviewers extracted data using a pre-piloted extraction form and strict inclusion criteria. Tool content was evaluated for domains potentially related to bias and was informed by the STROBE guidelines for reporting observational epidemiological studies.

Results A total of 86 tools were reviewed, comprising 41 simple checklists, 12 checklists with additional summary judgements and 33 scales. The number of items ranged from 3 to 36 (mean 13.7). One-third of tools were designed for single use in a specific review and one-third for critical appraisal. Half of the tools provided development details, although most were proposed for future use in other contexts. Most tools included items for selection methods (92%), measurement of study variables (86%), design-specific sources of bias (86%), control of confounding (78%) and use of statistics (78%); only 4% addressed conflict of interest. The distribution and weighting of domains across tools was variable and inconsistent.

Conclusion A number of useful assessment tools have been identified by this report. Tools should be rigorously developed, evidence-based, valid, reliable and easy to use. There is a need to agree on critical elements for assessing susceptibility to bias in observational epidemiology and to develop appropriate evaluation tools.

Keywords Observational studies, epidemiological studies, quality, bias, checklist, scales

Introduction

Systematic reviews identify, appraise and synthesize evidence from multiple studies of the same research question, and can be applied to diverse topics in medical research, including the effects of health-care interventions, the accuracy of diagnostic tests and the relationship between risk factors and disease. Meta-analyses, often contained within systematic reviews, offer a means of quantitatively summarizing the body of evidence identified. The strengths and limitations of systematic reviews and meta-analyses have been well established for randomized clinical trials, largely through the efforts of The Cochrane Collaboration. Although they have been used in parallel for observational epidemiological studies, such as cohort, case-control and cross-sectional studies, considerably less attention has been paid to their methodology in this area of application.

A systematic review should follow a protocol in order to minimize bias and ensure that the findings are reproducible.

[1] Primary Care Genetics, General Practice and Primary Care Research Unit, University of Cambridge and Public Health Genetics Unit, Cambridge,

[2] Public Health Genetics Unit, Cambridge.

[3] MRC Biostatistics Unit, Cambridge and Public Health Genetics Unit, Cambridge, UK.

[4] Present address: PBSE, Hoffman-La Roche, Basel, Switzerland.

* Corresponding author and guarantor. Strangeways Research Labs, Worts Causeway, Cambridge CB1 8RN, UK.
E-mail: simon.sanderson@srl.cam.ac.uk

A key source of potential bias in a meta-analysis is bias due to limitations in the original studies contained within it. For example, a review of case-control studies of oral contraceptives and risk of rheumatoid arthritis found exaggerated effects in hospital-based control groups compared with population-based control groups[1] whilst a review of case-control studies investigating the impact of sunlight exposure on skin cancer identified an important difference between study results when subjects or interviewers were blinded (or not) to skin cancer status.[2] A large prospective study of the association between C-reactive protein and coronary heart disease obtained odds ratios varying from 2.13 to 3.46 with different degrees of adjustment for confounding variables.[3]

An important component of a thorough systematic review is therefore an evaluation of the methodological quality of the primary research. Numerous tools have been proposed for evaluation of methodological quality of observational epidemiological studies. A comprehensive study of tools for assessing non-randomized intervention studies in health care (excluding case-control studies) identified 193 tools, including several that could also be used for assessing non-intervention studies.[4] A large-scale review of tools for grading the quality of research articles and rating the strength of bodies of evidence identified 17 tools for grading evidence from observational study designs,[5] although it did not include some of the key tools identified in previous reviews. More recently, Katrak and colleagues[6] reviewed 121 critical appraisal tools for allied health research, including physiotherapy, occupational and speech therapy and found a number of problems. All of these reviews have generally concluded that there is currently no agreed 'gold standard' appraisal tool; that the majority of tools did not undergo a rigorous development process; and that there are many tools from which to choose. Consequently, to our knowledge, no tool has been adopted for widespread use within systematic reviews. In addition, none of these reviews sought to identify all tools for assessing observational epidemiological studies.

'Quality' is an amorphous concept. A convenient interpretation is 'susceptibility to bias', although it is not uncommon for aspects of study conduct that are not directly associated with bias to be included in a quality assessment. For example, study size, whether or not a power calculation was performed, and ethical approval might be considered aspects of quality, but are, in their own right, not potential causes of bias. Our main objective was to seek tools to assess susceptibility to bias, but we do not draw a clear distinction between quality in bias, reflecting the lack of a distinction in much of the published literature.

It is important, however, to distinguish between quality of reporting and quality of what was actually done in the design, conduct and analysis of a study. A high-quality report ensures that all relevant information about a study is available to the reader, but does not necessarily reflect a low susceptibility to bias.[1] Factors such as the peer-review process, editorial policy or journal space restrictions may preclude detailed reporting and so make it difficult to assess inherent biases. A number of consensus statements have encouraged higher quality of reporting, including recommendations for reporting systematic reviews (QUOROM),[7] randomized trials (CONSORT),[8] studies of diagnostic tests (STARD),[9] meta-analyses of observational studies (MOOSE)[10] and observational epidemiological studies (STROBE).[11,12] These are aimed at authors of reports, not at those seeking to assess the validity of what they read.

This study provides an annotated bibliography of tools specifically designed to assess quality or susceptibility to bias in observational epidemiological studies, obtained from a comprehensive search of the published literature and of the Internet. It follows the approach of a previous review of tools to assess quality of randomized controlled trials,[13] and attempts to identify whether there is an existing tool that could be recommended for widespread use.

Methods

Inclusion criteria

To be included in the review, a tool was defined as any structured instrument aimed at aiding the user to assess quality or susceptibility to bias in observational epidemiological studies (cohort, case-control and cross-sectional studies). Tools were placed in one of the following three categories defined below: scales, simple checklists or checklists with a summary judgement. Scales result in a summary numerical score, typically derived as a sum of scores for several items. Checklists consisted of only a list of items, whilst checklists with a summary judgement were checklists that also resulted in an overall qualitative assessment about the study's quality, such as 'high', 'medium' or 'low'. These tools may have been developed for use in critical appraisal or in systematic reviews, and may have been developed for general use or use in a specific context. Articles that provided general narrative guidance only or were without an explicit scale or checklist were excluded.

Search methods

Three electronic databases (MEDLINE, EMBASE and Dissertation Abstracts up to March 2005) were searched using full text and MeSH terms to identify articles discussing observational epidemiological study designs, including 'cohort studies', 'case-control studies', 'cross-sectional studies' and 'follow-up studies'. Where possible, all terms were included as full text, with truncation used where possible to capture variation in the terminology. The search was not limited to the English language, nor restricted by any other means.

In order to capture tools posted on Internet websites, we conducted an Internet search using the Google® search engine[14] during March 2005. Searches were conducted using several combinations of the following search terms: 'tool', 'scale', 'checklist', 'validity', 'quality', 'critical appraisal', 'bias' and 'confounding'. The first 300 links identified by each separate search were investigated. Reference lists of published articles were examined to identify additional sources not identified in the database searches.

Inclusion criteria

Articles or websites were included if they described a tool suitable for assessing quality of observational epidemiological studies. Abstracts were scrutinized for suitability before obtaining the full text of all relevant articles. Where more than one tool was published within the same article or website (for example, independent tools for assessing cohort and case-control study designs published within the same article or website), these were included as separate quality assessment tools. Published reports were used in preference to web sites for tools reported in both formats. Care was taken not to include the same tool twice.

Data extraction

A data extraction form was developed and piloted and included information about the type of study addressed by the tool, number of items, scoring system, description of the development process, whether the tool was developed for generic use in systematic reviews, single use in a specific systematic review or for critical appraisal, and whether the tool was proposed for future use. Data extraction was performed by two authors (SS and IT) with differences of opinion resolved by discussion or by the third author (JH). Items in tools were classified into domains that covered key potential sources of bias. The selection was strongly influenced by the 'STrengthening the Reporting of OBservational studies in Epidemiology' (STROBE) guidelines for reporting observational epidemiological studies. These guidelines for reporting case-control, cohort and cross-sectional studies were developed by an international collaboration of epidemiologists, statisticians and journal editors. Although not a tool for assessing the quality of primary studies, they provide a useful indication of the essential information needed to appraise the conduct of such studies. Table 1 shows how the domains and criteria were used to evaluate tool content.

Wherever possible, we have attempted to demonstrate weighting within checklists and scales by including the total number of items for a checklist and the number of these items allocated to a particular quality domain. For scales, we have included the total maximum raw score for each scale and the possible total score by domain (although most scales do not address all of the domains in Table 1). A few of the tools use extremely complicated assessment and scoring systems, and for these we have reported the total raw score and the maximum item score by domain.

Results

A total of 86 tools were included in the review, 62 identified from the electronic database search (72%) and a further 24 from the Internet search (28%). An overall summary of the main tool characteristics is presented in Tables 2–4 and more detailed information in Tables 5–7.

The biggest group was checklists (41; 48%),[15–46] followed by scales (33; 38%)[47–73] and finally summary judgement checklists (12; 14%)[74–82]. Fifteen per cent of all tools were for generic use in systematic reviews, one-third for use in critical appraisal, one-third for single use in a specific systematic review and 15% where the purpose was ambiguous. For checklists, half were critical appraisal tools (22; 54%) whilst two-thirds of scales were review-specific (21; 64%). Over half of all tools (54%) described their development process in detail.

Just under three-quarters of all tools were proposed as being suitable for future use, including all of the critical appraisal tools and generic systematic review tools and six of the tools originally designed for use in a specific systematic review.

A number of tools were designed to address specific study design types: case-control studies alone (19%); cohort studies alone (27%) and cross-sectional studies alone (7%) (Table 3). Others addressed different combinations of these design types, with almost one-third addressing both case-control and cohort studies (45%) and 15% addressing all three. The number of items in all tools ranged from 3 to 36, with a mean of 13.7 (13.4 for simple checklists, 15.2 for simple checklists with a summary judgement and 12.6 for scales).

The majority of tools included items relating to methods for selecting study participants (92%). The proportion of tools including items about the measurement of study variables (exposure, outcome and/or confounding variables) was also high (86%). Assessment of other design-specific sources of bias (including recall bias, interviewer bias and biased loss to follow-up but excluding confounding) was included in 86%, around three-quarters assessed control of confounding (78%) and three-quarters included items concerning statistical methods (78%). Conflict of interest was included in only three tools (3%).

To address weighting, we recorded the number of items included in both types of checklists devoted to each of our key domains, whilst for scales we recorded the total available raw score for each domain. As can be seen from Tables 5 to 7, there is a little consistency among tools, with considerable variability in the number of items across domains and across tool types.

Table 1 Domains and criteria for evaluating each tool's content

Domain	Tool item must address
Methods for selecting study participants	Appropriate source population (cases, controls and cohorts) ***and*** inclusion ***or*** exclusion criteria
Methods for measuring exposure and outcome variables	Appropriate measurement methods for ***both*** exposure(s) and/or outcome(s)
Design-specific sources of bias (excluding confounding)	Appropriate methods outlined to deal with any design-specific issues such as recall bias, interviewer bias, biased loss to follow or blinding
Methods to control confounding	Appropriate design ***and/or*** analytical methods
Statistical methods (excluding control of confounding)	Appropriate use of statistics for primary analysis of effect
Conflict of interest	Declarations of conflict of interest **or** identification of funding sources

Table 2 Summary results comparing identified tools by type

Tool characteristics	Simple checklists ($n=41$)	Simple checklists with additional judgement ($n=12$)	Scales ($n=33$)	Total ($n=86$)
Source				
Electronic database	21 (51%)[a]	9 (75%)	32 (97%)	**62 (72%)**
Internet	20 (49%)	3 (25%)	1 (3%)	**24 (28%)**
	100%	*100%*	*100%*	***100%***
Tool purpose				
Single use in a specific context	3 (7%)	4 (33%)	22 (67%)	**29 (34%)**
Generic tool for systematic reviews	8 (20%)	3 (25%)	2 (6%)	**13 (15%)**
Critical appraisal tool	22 (54%)	4 (33%)	5 (15%)	**31 (36%)**
Ambiguous (unable to allocate above categories)	8 (20%)	1 (8%)	4 (12%)	**13 (15%)**
	41 (100%)	*12 (100%)*	*33 (100%)*	***86 (100%)***
Development				
Development described	21 (51%)	7 (58%)	18 (55%)	**46 (53%)**
Future use				
Proposed for future use	38 (93%)	8 (67%)	14 (42%)	**60 (70%)**

[a] Percentages subject to rounding error.

Table 3 Summary results comparing identified tools by content

	Simple checklists ($n=41$)	Simple checklists with additional judgement ($n=12$)	Scales ($n=33$)	Total ($n=86$)
Tool content				
Number of items				
• Range	3–36	4–32	4–35	
• Mean	13.4	15.2	12.6	
Maximum raw score range (scales only)	NA	NA	4–72	
Appropriate methods for selecting study participants % (range)	39; 95%[a] (1–10)	11; 92% (1–6)	29; 88% (1–26.4)	**79 (92%)**
Appropriate methods for measuring exposure and outcome variables % (range)	36; 88% (1–10)	12; 100% (1–8)	26; 79% (1–22)	**74 (86%)**
Appropriate design-specific sources of bias (excluding confounding) *n*; % (range)	36; 88% (1–6)	11; 92% (1–10)	27; 82% (1–8)	**74 (86%)**
Appropriate methods to control confounding *n*; % (range)	34; 83% (1–5)	12; 100% (1–3)	21; 64% (1–12)	**67 (78%)**
Appropriate statistical methods (primary analysis of effect but excluding confounding) *n*; % (range)	34; 83% (1–8)	8; 67% (1–3)	24; 73% (1–20)	**66 (78%)**
Conflict of interest *n*; % (range)	1; 2% (1)	1; 8% (1)	1; 3% (1)	**3 (4%)**

Note: For checklists, the range represents items; for scales, it represents available raw scores.
[a] Percentages subject to rounding error.

Table 4 Distribution of tools by epidemiological study design addressed

Case-control	Cohort	Cross-sectional	Simple checklists *n* (%)	Simple checklists with a judgement *n* (%)	Scales *n* (%)	Total *n* (%)
Y	N	N	9 (22)	2 (17)	5 (15)	**16 (19)**
Y	Y	N	15 (36)	6 (50)	7 (21)	**28 (32)**
Y	Y	Y	4 (10)	1 (8)	8 (24)	**13 (15)**
N	Y	N	11 (27)	2 (17)	10 (30)	**23 (27)**
N	N	Y	2 (5)	1 (8)	3 (9)	**6 (7)**
			41	**12**	**33**	**86**

Note: Y = yes; N = no.

Table 5 Simple checklists

Study/tool name/reference ID	Year	Source	Tool purpose	CC	Coh	CS	Items (*n*)	Development described	Future use	Participants	Variables measure	Other biases	Control confounding	Other statistics	Conflict of interest
Avis[15]	1994	ED	CA	Y	Y		24	Y	Y	5	N	2	1	1	N
Briggs[16]	@	W	AMB	Y	Y	Y	5	N	N	1	1	N	1	1	N
Cameron[17]	2000	ED	SU	Y	Y		36	N	Y	4	3	6	1	2	N
Carneiro[18]	2002	ED	CA		Y		8	N	Y	1	1	1	1	1	N
CASP CC[19]	@	W	CA	Y			7	N	Y	2	1	1	1	1	N
CASP Co[19]	@	W	CA		Y		8	N	Y	1	2	2	2	1	N
CenOccHealth[20]	@	W	CA	Y	Y	Y	23	N	Y	8	10	2	N	2	Y
CEBM Prog[21]	@	W	CA		Y		7	Y	Y	1	N	2	1	1	N
CEBM Diag[21]	@	W	CA	Y	Y	Y	3	Y	Y	1	1	1	N	N	N
DuRantCC[22]	1994	ED	CA	Y			22	N	Y	6	4	5	3	3	N
DuRantCoh[22]	1994	ED	CA		Y		24	N	Y	7	8	2	2	3	N
DuRantCS[22]	1994	ED	CA			Y	18	N	Y	6	4	2	2	3	N
Elwood[23]	2002	ED	CA	Y	Y		20	Y	Y	Y	2	1	1	1	N
Esdaile[24]	1985	ED	SU	Y	Y		6	Y	N	N	2	1	1	N	N
Gardner[25]	1986	ED	AMB	Y	Y		12	N	Y	1	N	N	N	7	N
Hadorn[26]	1996	ED	AMB		Y		24	Y	Y	8	3	2	1	8	N
HEB Wales[27]	@	W	CA	Y	Y	Y	13	Y	Y	2	1	4	3	1	N
Horwitz[28]	1979	ED	CA	Y			12	N	Y	6	2	2	N	N	N
Khan[29]	@	W	SR	Y			9	N	Y	2	2	2	1	2	N
Khan[29]	@	W	SR		Y		10	Y	Y	2	1	4	2	1	N
Kilgore[30]	1981	ED	CA	Y	Y		2	Y	Y	N	N	N	N	N	N
Levine[31]	1994	ED	CA	Y	Y		7	N	Y	1	1	1	1	2	N
Lichtenstein[32]	1987	ED	CA	Y			20	Y	Y	4	2	4	2	3	N
London[33]	@	W	CA	Y	Y		30	Y	Y	4	10	5	5	3	N
Margetts[34]	2002	ED	SR	Y	Y		6	N	Y	2	2	1	N	2	N
Montreal[35]	@	W	CA	Y	Y		8	N	Y	2	1	1	1	1	N
Mulrow[36]	1986	ED	SU		Y		9	Y	N	2	2	2	1	1	N
Newc-Ott CC[37]	@	W	SR	Y			8	N	Y	4	2	1	1	N	N
Newc-Ott Co[37]	@	W	SR		Y		8	N	Y	2	3	2	1	N	N
QUADAS[38]	2003	ED	SR	Y	Y		14	Y	Y	2	3	3	N	N	N
Campbell[39]	2003	ED	AMB	Y			13	N	Y	Y	Y	N	Y	Y	N
SIGN 50 CC[40]	@	W	AMB	Y			22	Y	Y	6	1	2	1	3	N
SIGN 50 Co[40]	@	W	AMB		Y		25	Y	Y	5	3	4	1	3	N
Solomon[41]	1997	ED	SR	Y	Y		12	N	Y	1	3	1	2	1	N

(continued)

Table 5 Continued

Study/tool name/reference ID	Year	Source	Tool purpose	CC	Coh	CS	Items (*n*)	Development described	Future use	Participants	Variables measure	Other biases	Control confounding	Other statistics	Conflict of interest
STARD[42]	@	W	AMB	Y	Y		14	N	Y	3	4	3	1	2	N
Surgical tutor[43]	@	W	CA	Y	Y		18	Y	Y	Y	4	2	3	3	N
UCW CC[44]	@	W	CA	Y			6	Y	Y	1	2	1	2	1	N
UCW Co[44]	@	W	CA		Y		8	Y	Y	1	N	4	1	1	N
UCW Cross[44]	@	W	CA			Y	3	Y	Y	1	N	N	1	1	N
Zaza[45]	2000	ED	SR	Y	Y		15	Y	Y	5	3	3	1	2	N
Zola[46]	1989	ED	AMB		Y		11	Y	Y	2	2	2	N	2	N

Note: ED, electronic database; W, Internet search; CA, critical appraisal; SR, for conducting systematic reviews; SU, single use in specific context; AMB, ambiguous; these purpose of these tools was not easy to determine and they could be designed for use in guideline development, reporting, critically appraising and/or integrating study data; CC, case-control; Coh, cohort; CS, cross-sectional; NA, not available; NR, not recorded; @, accessed during March 2005; Y, item addressed relevant domain and/or raw score or number of items unavailable; N, domain not addressed.

Table 6 Checklists with an additional summary judgement

Study/tool name/ reference ID	Year	Source	Purpose	CC	Coh	CS	Items (*n*)	Development described	Future use	Participants	Variables measure	Other biases	Control confounding	Other statistics	Conflict of interest
Bollini[74]	1992	ED	SU	Y	Y		10	Y	N	3	3	1	2	N	N
Ciliska[75]	1996	ED	SU	Y	Y		6	Y	N	N	1	1	1	1	N
Cowley[76]	1995	ED	SU	Y	Y		13	N	N	1	1	3	1	2	1
Effective PH[77]	@	W	CA	Y	Y		13	N	Y	2	2	2	3	3	N
EPIQ CC[78]	@	W	CA	Y			30	N	Y	5	8	8	3	3	N
EPIQ Cohort[78]	@	W	CA		Y		32	N	Y	6	8	10	3	3	N
Fowkes[79]	1991	ED	CA	Y	Y	Y	22	N	Y	6	3	6	2	2	N
GyorkosCC[80]	1994	ED	SR	Y			5	Y	Y	2	1	1	1	N	N
GyorkosCoh[80]	1994	ED	SR		Y		6	Y	Y	1	2	2	1	N	N
GyorkosCS[80]	1994	ED	SR			Y	4	Y	Y	1	2	N	1	N	N
Spitzer[81]	1990	ED	SU	Y	Y		17	Y	N	4	4	3	3	2	N
Steinberg[82]	2000	ED	AMB	Y	Y		24	Y	Y	3	2	5	2	3	N

Note: ED, electronic database; W, Internet search; CA, critical appraisal; SR, for conducting systematic reviews; SU, single use in specific context; AMB, ambiguous; these purpose of these tools was not easy to determine and they could be designed for use in guideline development, reporting, critically appraising and/or integrating study data; CC, case-control; Coh, cohort; CS, cross-sectional; NA, not available; NR, not recorded; @, accessed during March 2005; Y, item addressed relevant domain and/or raw score or number of items unavailable; N, domain not addressed.

Table 7 Scales

Study/tool name/ reference ID	Year	Source	Purpose	CC	Coh	CS	Items (*n*)	Development described	Future use	Maximum raw score	Participants	Variables measure	Other biases	Control confounding	Other statistics	Conflict of interest
Anders[47]	1996	ED	SU		Y		6	N	N	6	N	3	2	N	N	N
AriensCC[48]	2000	ED	SU	Y			18	Y	Y	18	3	9	1	1	2	N
AriensCoh[48]	2000	ED	SU		Y		17	Y	Y	17	3	7	1	1	2	N
AriensCS[48]	2000	ED	SU			Y	13	Y	Y	13	2	8	1	1	2	N
Berlin[49]	1990	ED	SU	Y	Y		16	Y	N	32	N	2	N	2	N	N
Bhutta[50]	2002	ED	SU	Y			6	N	Y	10	1	N	N	2	N	N
Borghouts[51]	1998	ED	SU		Y		13	Y	N	13	3	1	1	N	2	N
Campos[52]	1995	ED	SU	Y	Y		7	N	N	70	N	10	N	N	10	N
Carson[53]	1994	ED	AMB		Y		10	Y	Y	10	3	N	2	1	2	N
Loney[54]	@	W	CA	Y	Y	Y	6	N	Y	8	2	2	1	N	1	N
Cho[55,b]	1994	ED	CA	Y	Y	Y	18	Y	Y	36	12	2	8	4	2	N
Corrao[56]	1999	ED	SU	Y	Y		16	N	N	30	5	9	4	N	2	N
Downs[57]	1998	ED	CA	Y	Y		17	Y	Y	21	5	1	3	3	8	N
Garber[68]	1996	ED	SU	Y	Y	Y	6	N	N	18	N	N	N	N	N	N
Goodman[59]	1994	ED	AMB	Y	Y		10	Y	Y	50	20	N	5	5	15	N
Jabbour[60]	1996	ED	SU		Y		7	N	N	7	1	N	1	N	N	N
Kreulen[61,c]	1998	ED	SU			Y	16	N	N	42	3	12	6	3	12	N
Krogh[62]	1985	ED	CA	Y	Y	Y	7	N	Y	4	2	N	1	N	1	N
Littenberg[63,d]	1998	ED	SU	Y	Y	Y	15	N	N	45	NA	NA	NA	NA	NA	N
LongneckerCC[64,a]	1988	ED	SU	Y			11	N	N	53/58[a]	(5)	(5)	(5)	(5)	N	N
LongneckerCoh[64]	1988	ED	SU		Y		4	N	N	20	5	5	5	5	N	N
Macfarlane[65]	2001	ED	AMB	Y	Y	Y	6	Y	N	6	2	1	2	N	1	N
Manchikanti[66]	2002	ED	SU	Y	Y		6	Y	N	6	2	2	N	N	1	1
MargettsCC[67,a]	1995	ED	SR	Y			13	Y	Y	46.4	26.4	10	2	5	5	N

(continued)

Table 7 Continued

Study/tool name/ reference ID	Year	Source	Purpose	CC	Coh	CS	Items (*n*)	Development described	Future use	Maximum raw score	Participants	Variables measure	Other biases	Control confounding	Other statistics	Conflict of interest
MargettsCoh[67]	1995	ED	SR		Y		19	Y	Y	53.4	4	22	7	2	9	N
Meijer[68]	2003	ED	SU		Y		9	Y	Y	9	1	3	1	1	2	N
Nguyen[69]	1999	ED	SU	Y	Y	Y	14	N	N	72	4	18	6	12	20	N
Rangel[70]	2003	ED	AMB		Y		15	Y	N	17	3	5	1	N	6	N
Reisch[71,a]	1989	ED	CA	Y	Y		35 (min)	N	Y	% items fulfilled	(1)	(1)	(1)	N	(1)	N
Stock[72]	1991	ED	SU	Y	Y	Y	7	N	N	21	6	6	3	3	N	N
WindtCC[73]	2000	ED	SU	Y			20	Y	N	20	3	11	4	1	4	N
WindtCoh[73]	2000	ED	SU		Y		18	Y	N	18	3	8	1	1	4	N
WindtCS[73]	2000	ED	SU			Y	16	Y	N	16	3	8	1	1	4	N

Note: ED, electronic database; W, Internet search; CA, critical appraisal; SR, For conducting systematic reviews; SU, Single use in specific context; AMB, Ambiguous; these purpose of these tools was not easy to determine and they could be designed for use in guideline development, reporting, critically appraising and/or integrating study data; CC, case-control; Coh, cohort; CS, cross-sectional; NA, not available; NR, not recorded; @, accessed during March 2005; Y, item addressed relevant domain and/or raw score or number of items unavailable; N, domain not addressed.

[a] These tools were extremely complex and require considerable input to calculate raw scores and to convert to final scores, depending on the primary study design and methods.

[b] This tool allowed the possibility of different total scores based on study design and applied differential weighting, and included case studies and randomized trials within a single scale.

[c] Weighting was applied to the raw scores by a factor of 2 for study methodology, evaluation methodology and by a factor of 1.5 for statistical methodology.

[d] The scale is not described in sufficient detail to assess weighting in domains.

Discussion

Assessing the quality of evidence from observational epidemiological studies requires tools that are designed and developed with this specific purpose in mind. To our knowledge, this is the most comprehensive search to date of both the medical literature and the Internet for tools to assess such studies. We have identified 86 candidate tools, comprising checklists, summary judgement checklists and scales. The Internet search identified three more tools that were not identified through searching electronic databases. Future search strategies may wish to employ similar methodologies to ensure the identification of all available tools, articles or studies. Despite the comprehensive nature of the search strategy employed, it is unlikely that all existing tools for assessing quality of observational epidemiological studies have been identified, since many are developed for specific systematic reviews, and it is very difficult to identify all of these through searching electronic databases.

A large number of the tools were scales that resulted in numerical summary scores. Whilst this approach has the appearance of simplicity, considerable concerns have been raised about such an approach to assessing quality.[83] Summary scores involve inherent weighting of component items, some of which may not be directly related to the validity of a study's findings (such as sample size calculations). It is unclear how weights for different items should be determined, and different scales may reach different conclusions on the overall quality of an individual study.[84] We have found that the weighting applied in scales to different study domains is variable and inconsistent. Similar considerations apply to summary judgement checklists, although qualitative rather than quantitative summaries may be less prone to inappropriate analysis. We prefer a more transparent checklist approach that concentrates on the few, principal, potential sources of bias in a study's findings.

Tool components should, where possible, be based on empirical evidence of bias, although this may be difficult to obtain, and there is a need for more empirical research on relationships between specific quality items and findings from epidemiological studies. There was wide variation among tools in the number and nature of items, scoring ranges (where applicable) and levels of development. The specific components assessed by the tools differed across both study design and tool type. Although we have not implemented all tools, we would anticipate that different tools would indicate different degrees of quality when applied to the same study.

It is encouraging that most tools included items to assess methods for selecting study participants (92%) and to assess methods for measuring study variable and design-specific sources of bias (both 86%). Over three-quarters of tools assessed the appropriate use of statistics, and the control of confounding (both 78%) but conflict of interest was only included in 4% of tools. Around one-third of the tools were designed for specific clinical or research topics, limiting their wider applicability; there was a marked difference between tool types in this respect, with the majority of checklists designed for critical appraisal and the majority of scales for single use in specific single reviews. The ambiguity of purpose of some of the tools is a cause for concern,

and more clarity is needed to differentiate assessments of the quality of reporting from the quality of what was actually done in the study.

A rigorous development process should be an important component of tool design, but only half of the tools provided a clear description of their design, development or the empirical basis for item inclusion or evaluation of the tool's validity and reliability. This is of particular concern as 70% of the tools were proposed as being suitable for future use in other contexts. Future tools should undergo a rigorous development process to ensure that they are evidence-based, easy to use and readily interpretable.

This review has highlighted the lack of a single obvious candidate tool for assessing quality of observational epidemiological studies. One might regard this review as the first stage towards development of a generic tool. In such an endeavour, one would need to reach a consensus on the critical domains that should be included. The development of the STROBE statement has involved extensive discussion among numerous experienced epidemiologists and statisticians. Despite targeting the reporting of studies, many items were no doubt selected due to presumed (or evidence of) association with susceptibility to bias. Thus the statement should provide a suitable starting point for development of a quality assessment tool, and we have been guided by it in our presentation of results.

Around half of the checklists included what we regard as the three most fundamental domains of appropriate selection of participants, appropriate measurement of variables and appropriate control of confounding; all were considered appropriate for future use. The majority of these tools also included items on potential design-specific biases. However, we are reluctant to recommend a specific tool, without having implemented them all on multiple studies with a view to assessing their properties and ease-of-use. Our broad recommendations are that tools should (i) include a small number of key domains; (ii) be as specific as possible (with due consideration of the particular study design and topic area); (iii) be a simple checklist rather than a scale and (iv) show evidence of careful development, and of their validity and reliability.

Search strategy

- scale*
- checklist*
- critical apprais*
- tool*
- valid*
- quality
- (bias* OR confounding) AND (assess* OR measure* OR evaluat*)
- OBSERVATIONAL STUDIES (MeSH)
- observational stud*
- COHORT STUDIES (MeSH)
- cohort stud*
- CASE-CONTROL STUDIES (MeSH)
- case-control stud*
- CROSS-SECTIONAL STUDIES (MeSH)
- cross-sectional stud*
- FOLLOW-UP STUDIES (MeSH)
- follow-up stud*

(1 or 2 or 3 or 4) AND (5 or 6 or 7) AND (8 or . . . to 17)

Conflict of interest: None declared.

KEY MESSAGES

- Tools for assessing quality in clinical trials are well-described but much less attention has been given to similar tools for observational epidemiological studies.
- Only about half of the identified tools did not describe their development or validity and reliability.
- Tools for assessing quality should be rigorously developed, evidence-based, valid, reliable and easy to use and concentrate on assessing sources of bias.
- There is a need to agree on critical elements for assessing susceptibility to bias in observational epidemiology and to develop appropriate evaluation tools.

References

[1] Huwiler-Muntener K, Juni P, Junker C, Egger M. Quality of reporting of randomized trials as a measure of methodologic quality. *JAMA* 2002;**287**:2801–4.

[2] Nelemans PJ, Rampen FH, Ruiter DJ, Verbeek AL. An addition to the controversy on sunlight exposure and melanoma risk: a meta-analytical approach. *J Clin Epidemiol* 1995;**48**:1331–42.

[3] Danesh J, Whincup P, Walker M *et al*. Low grade inflammation and coronary heart disease: prospective study and updated meta-analyses. *BMJ* 2000;**321**:199–204.

[4] Pladevall-Vila M, Delclos GL, Varas C, Guyer H, Brugues-Tarradellas J, Anglada-Arisa A. Controversy of oral contraceptives and risk of rheumatoid arthritis: meta-analysis of conflicting studies and review of conflicting meta-analyses with special emphasis on analysis of heterogeneity. *Am J Epidemiol* 1996;**144**:1–14.

[5] Juni P, Altman DG, Egger M. Systematic reviews in health care: assessing the quality of controlled clinical trials. *BMJ* 2001;**323**:42–46.

[6] Katrak P, Bialocerkowski AE, Massy-Westropp N, Kumar S, Grimmer KA. A systematic review of the content of critical appraisal tools. *BMC Med Res Methodol* 2004;**4**:22.

[7] Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Quality of reporting of meta-analyses. *Lancet* 1999;**354**:1896–900.
[8] Deeks JJ, Dinnes J, D'Amico R *et al*. Evaluating non-randomised intervention studies. *Health Technol Assess* 2003;**7**:iii–173.
[9] West S, King V, Carey TS, Lohr KN, McKoy N, Sutton SF, Lux L. Systems to Rate the Strength of Evidence. Evidence Report/ Technology Assessment No. 47. 2002. Agency for Healthcare Research and Quality, Rockville, MD. AHRQ Publication No. 02-E016.
[10] Stroup DF, Berlin JA, Morton SC *et al*. Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. *JAMA* 2000;**283**:2008–12.
[11] von Elm E, Egger M. The scandal of poor epidemiological research. *BMJ* 2004;**329**:868–69.
[12] Altman D, Egger M, Pocock S, Vandenbrouke JP, von Elm E. Strengthening the reporting of observational epidemiological studies. STROBE Statement: Checklist of Essential Items Version 3 (September 2005) http://www.strobe-statement.org/Checkliste.html).
[13] Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. *Control Clin Trials* 1995;**16**:62–73.
[14] Google Home page. *Google Home page*. 2004.
[15] Avis M. Reading research critically. II. An introduction to appraisal: assessing the evidence. *J Clin Nurs* 1994;**3**:271–77.
[16] The Joanna Briggs Institute. System for the Unified Management of the Review and Assessment of Information (SUMARI). The Joanna Briggs Institute 2004.
[17] Cameron I, Crotty M, Currie C *et al*. Geriatric rehabilitation following fractures in older people: a systematic review. *Health Technol Assess* 2000;**4**:i–111.
[18] Carneiro AV. Critical appraisal of prognostic evidence: practical rules. *Rev Port Cardiol* 2002;**21**:891–900.
[19] CASP, NHS. Critical Appraisal Skills Programme (CASP): appraisal tools. Public Health Resource Unit, NHS 2003.
[20] Centre for Occupational and Environmental Health. Critical Appraisal. School of Epidemiology and Health Sciences, University of Manchester 2003.
[21] Centre for Evidence-Based Mental Health, University of Oxford. Critical Appraisal Forms. http://www.cebmh.com/, 2004.
[22] DuRant RH. Checklist for the evaluation of research articles. *J Adolesc Health* 1994;**15**:4–8.
[23] Elwood M. Forward projection—using critical appraisal in the design of studies. *Int J Epidemiol* 2002;**31**:1071–73.
[24] Esdaile JM, Horwitz RI. Observational studies of cause-effect relationships: an analysis of methodologic problems as illustrated by the conflicting data for the role of oral contraceptives in the etiology of rheumatoid arthritis. *J Chronic Dis* 1986; **39**:841–52.
[25] Gardner MJ, Machin D, Campbell MJ. Use of check lists in assessing the statistical content of medical studies. *Br Med J (Clin Res Ed)* 1986;**292**:810–12.
[26] Hadorn DC, Baker D, Hodges JS, Hicks N. Rating the quality of evidence for clinical practice guidelines. *J Clin Epidemiol* 1996;**49**:749–54.
[27] Health Evidence Bulletin, Wales. Questions to assist with the critical appraisal of an observational study eg cohort, case-control, cross-sectional. HEB, Wales, 2004.
[28] Horwitz RI, Feinstein AR. Methodologic standards and contradictory results in case-control research. *Am J Med* 1979;**66**:556–64.
[29] Khan KS, Riet GT, Popay J, Nixon J, Kleijnen J. Undertaking systematic reviews of research effectiveness. CRD's guidance for those carrying out or commissioning reviews. CRD Report number 4 (2nd edn). 2001. The University of York Centre for Reviews and Dissemination.
[30] Department of Clinical Epidemiology and Biostatistics. How to read clinical journals: IV. To determine etiology or causation. *Can Med Assoc J* 1981;**124**:985–90.
[31] Levine M, Walter S, Lee H, Haines T, Holbrook A, Moyer V. Users' guides to the medical literature. IV. How to use an article about harm. Evidence-Based Medicine Working Group. *JAMA* 1994;**271**:1615–19.
[32] Lichtenstein MJ, Mulrow CD, Elwood PC. Guidelines for reading case-control studies. *J Chronic Dis* 1987;**40**:893–903.
[33] Federal Focus, Incorporated. The London Principles for Evaluating Epidemiologic Data in Regulatory Risk Assessment. http://www.fedfocus.org/science/london-principles.html, 2004.
[34] Margetts BM, Vorster HH, Venter CS. Evidence-based nutrition—review of nutritional epidemiological studies. *South African J Clin Nutr* 2002;**15**:68–73.
[35] University of Montreal. Critical Appraisal Worksheet. University of Montreal, 2004.
[36] Mulrow CD, Lichtenstein MJ. Blood glucose and diabetic retinopathy: a critical appraisal of new evidence. *J Gen Intern Med* 1986;**1**:73–77.
[37] Wells GA, Shea B, O'Connell D, Peterson J, Welch V, Losos M, Tugwell P. Quality Assessment Scales for Observational Studies. Ottawa Health Research Institute, 2004.
[38] Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Med Res Methodol* 2003;**3**:25.
[39] Campbell H, Rudan I. Interpretation of genetic association studies in complex disease. *Pharmacogenomics J* 2002;**2**:349–60.
[40] Scottish Intercollegiate Guidelines Network. SIGN 50: A guideline developers' handbook. Scottish Intercollegiate Guidelines Network, 2004. Ref Type: Electronic Citation.
[41] Solomon DH, Bates DW, Panush RS, Katz JN. Costs, outcomes, and patient satisfaction by provider type for patients with rheumatic and musculoskeletal conditions: a critical review of the literature and proposed methodologic standards. *Ann Intern Med* 1997;**127**:52–60.
[42] The STARD Group. The STARD Initiative—Towards Complete and Accurate Reporting of Studies on Diagnostic Accuracy. http://www.consort-statement.org/stardstatement.htm, 2001.
[43] Critical appraisal: Guidelines for the critical appraisal of a paper. Surgical-Tutor.org.uk, 2004.
[44] University of Wales College of Medicine. Critical Appraisal Forms. University of Wales, 2004.
[45] Zaza S, Wright-De Aguero LK, Briss PA *et al*. Data collection instrument and procedure for systematic reviews in the guide to community preventive services. Task Force on Community Preventive Services. *Am J Prev Med* 2000;**18**:44–74.
[46] Zola P, Volpe T, Castelli G *et al*. Is the published literature a reliable guide for deciding between alternative treatments for patients with early cervical cancer? *Int J Radiat Oncol Biol Phys* 1989;**16**:785–97.
[47] Anders JF, Jacobson RM, Poland GA, Jacobsen SJ, Wollan PC. Secondary failure rates of measles vaccines: a meta-analysis of published studies. *Pediatr Infect Dis J* 1996;**15**:62–66.
[48] Ariens GA, van Mechelen W, Bongers PM, Bouter LM, van der WG. Physical risk factors for neck pain. *Scand J Work Environ Health* 2000;**26**:7–19.
[49] Berlin JA, Colditz GA. A meta-analysis of physical activity in the prevention of coronary heart disease. *Am J Epidemiol* 1990;**132**:612–28.
[50] Bhutta AT, Cleves MA, Casey PH, Cradock MM, Anand KJS. Cognitive and behavioral outcomes of school-aged children who were born preterm: a meta-analysis. *J Am Med Assoc* 2002;**288**: 728–37.
[51] Campos-Outcalt D, Senf J, Watkins AJ, Bastacky S. The effects of medical school curricula, faculty role models, and biomedical

research support on choice of generalist physician careers: a review and quality assessment of the literature. *Acad Med* 1995;**70**:611–19.
[52] Borghouts JA, Koes BW, Bouter LM. The clinical course and prognostic factors of non-specific neck pain: a systematic review. *Pain* 1998;**77**:1–13.
[53] Carson CA, Fine MJ, Smith MA, Weissfeld LA, Huber JT, Kapoor WN. Quality of published reports of the prognosis of community-acquired pneumonia. *J Gen Intern Med* 1994;**9**:13–19.
[54] Loney PL, Chambers LW, Bennett KJ, Roberts JG, Stratford PW. Critical appraisal of the health research literature: prevalence or incidence of a health problem. *Chronic Dis Canada* 2000;**19**:170–77.
[55] Cho MK, Bero LA. Instruments for assessing the quality of drug studies published in the medical literature. *JAMA* 1994;**272**: 101–4.
[56] Corrao G, Bagnardi V, Zambon A, Arico S. Exploring the dose-response relationship between alcohol consumption and the risk of several alcohol-related conditions: a meta-analysis. *Addiction* 1999;**94**:1551–73.
[57] Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *J Epidemiol Commun Health* 1998;**52**:377–84.
[58] Garber BG, Hebert PC, Yelle JD, Hodder RV, McGowan J. Adult respiratory distress syndrome: a systemic overview of incidence and risk factors. *Crit Care Med* 1996;**24**:687–95.
[59] Goodman SN, Berlin J, Fletcher SW, Fletcher RH. Manuscript quality before and after peer review and editing at Annals of Internal Medicine. *Ann Intern Med* 1994;**121**:11–21.
[60] Jabbour M, Osmond MH, Klassen TP. Life support courses: are they effective? *Ann Emerg Med* 1996;**28**:690–98.
[61] Kreulen CM, Creugers NH, Meijering AC. Meta-analysis of anterior veneer restorations in clinical studies. *J Dent* 1998;**26**:345–53.
[62] Krogh CL. A checklist system for critical review of medical literature. *Med Educ* 1985;**19**:392–95.
[63] Littenberg B, Weinstein LP, McCarren M *et al*. Closed fractures of the tibial shaft. A meta-analysis of three methods of treatment. *J Bone Joint Surg Am* 1998;**80**:174–83.
[64] Longnecker MP, Berlin JA, Orza MJ, Chalmers TC. A meta-analysis of alcohol consumption in relation to risk of breast cancer. *JAMA* 1988;**260**:652–56.
[65] Macfarlane TV, Glenny AM, Worthington HV. Systematic review of population-based epidemiological studies of oro-facial pain. *J Dent* 2001;**29**:451–67.
[66] Manchikanti L, Singh V, Vilims BD, Hansen HC, Schultz DM, Kloth DS. Medial branch neurotomy in management of chronic spinal pain: systematic review of the evidence. *Pain Physician* 2002;**5**:405–18.
[67] Margetts BM, Thompson RL, Key T *et al*. Development of a scoring system to judge the scientific quality of information from case-control and cohort studies of nutrition and disease. *Nutr Cancer* 1995;**24**:231–39.
[68] Meijer R, Ihnenfeldt DS, van Limbeek J, Vermeulen M, de Haan RJ. Prognostic factors in the subacute phase after stroke for the future residence after six months to one year. A systematic review of the literature. *Clin Rehabil* 2003;**17**:512–20.
[69] Nguyen QV, Bezemer PD, Habets L, Prahl-Andersen B. A systematic review of the relationship between overjet size and traumatic dental injuries. *Eur J Orthod* 1999;**21**:503–15.
[70] Rangel SJ, Kelsey J, Colby CE, Anderson J, Moss RL. Development of a quality assessment scale for retrospective clinical studies in pediatric surgery. *J Pediatr Surg* 2003;**38**:390–96.
[71] Reisch JS, Tyson JE, Mize SG. Aid to the evaluation of therapeutic studies. *Pediatrics* 1989;**84**:815–27.
[72] Stock SR. Workplace ergonomic factors and the development of musculoskeletal disorders of the neck and upper limbs: a meta-analysis. *Am J Ind Med* 1991;**19**:87–107.
[73] van der Windt DAWM, Thomas E, Pope DP *et al*. Occupational risk factors for shoulder pain: a systematic review. *Occup Environ Med* 2000;**57**:433–42.
[74] Bollini P, Garcia Rodriguez LA, Gutthann SP, Walker AM. The impact of research quality and study design on epidemiologic estimates of the effect of nonsteroidal anti-inflammatory drugs on upper gastrointestinal tract disease. *Arch Intern Med* 1992;**152**:1289–95.
[75] Ciliska D, Hayward S, Thomas H *et al*. A systematic overview of the effectiveness of home visiting as a delivery strategy for public health nursing interventions. *Can J Public Health* 1996;**87**:193–98.
[76] Cowley DE. Prostheses for primary total hip replacement. A critical appraisal of the literature. *Int J Technol Assess Health Care* 1995;**11**:770–78.
[77] Effective Public Health Practice Project. Quality Assessment Tool for Quantitative Studies. Effective Public Health Practice Project, 2003.
[78] School of Population Health. EPIQ (Effective Practice, Informatics and Quality Improvement). Faculty of Medical and Health Sciences, University of Auckland, 2004.
[79] Fowkes FG, Fulton PM. Critical appraisal of published research: introductory guidelines. *BMJ* 1991;**302**:1136–40.
[80] Gyorkos TW, Tannenbaum TN, Abrahamowicz M *et al*. An approach to the development of practice guidelines for community health interventions. *Can J Public Health* 1994;**85**:S8–S13.
[81] Spitzer WO, Lawrence V, Dales R *et al*. Links between passive smoking and disease: a best-evidence synthesis. A report of the Working Group on Passive Smoking. *Clin Invest Med* 1990;**13**: 17–42.
[82] Steinberg EP, Eknoyan G, Levin NW *et al*. Methods used to evaluate the quality of evidence underlying the National Kidney Foundation-Dialysis Outcomes Quality Initiative Clinical Practice Guidelines: description, findings, and implications. *Am J Kidney Dis* 2000;**36**:1–11.
[83] Greenland S, O'Rourke K. On the bias produced by quality scores in meta-analysis, and a hierarchical view of proposed solutions. *Biostatistics* 2001;**2**:463–67.
[84] Juni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA* 1999;**282**:1054–60.