Ma et al. Military Medical Research (2020) 7:7
https://doi.org/10.1186/s40779-020-00238-8

MMR MILITARY MEDICAL RESEARCH

REVIEW Open Access

Methodological quality (risk of bias) assessment tools for primary and secondary medical studies: what are they and which is better?

Lin-Lu Ma[1], Yun-Yun Wang[1,2], Zhi-Hua Yang[1], Di Huang[1,2], Hong Weng[1] and Xian-Tao Zeng[1,2,3,4*]

Abstract

Methodological quality (risk of bias) assessment is an important step before study initiation usage. Therefore, accurately judging study type is the first priority, and the choosing proper tool is also important. In this review, we introduced methodological quality assessment tools for randomized controlled trial (including individual and cluster), animal study, non-randomized interventional studies (including follow-up study, controlled before-and-after study, before-after/ pre-post study, uncontrolled longitudinal study, interrupted time series study), cohort study, case-control study, cross-sectional study (including analytical and descriptive), observational case series and case reports, comparative effectiveness research, diagnostic study, health economic evaluation, prediction study (including predictor finding study, prediction model impact study, prognostic prediction model study), qualitative study, outcome measurement instruments (including patient - reported outcome measure development, content validity, structural validity, internal consistency, cross-cultural validity/ measurement invariance, reliability, measurement error, criterion validity, hypotheses testing for construct validity, and responsiveness), systematic review and meta-analysis, and clinical practice guideline. The readers of our review can distinguish the types of medical studies and choose appropriate tools. In one word, comprehensively mastering relevant knowledge and implementing more practices are basic requirements for correctly assessing the methodological quality.

Keywords: Methodological quality, Risk of bias, Quality assessment, Critical appraisal, Methodology checklist, Appraisal tool, Observational study, Qualitative study, Interventional study, Outcome measurement instrument

Background

In the twentieth century, pioneering works by distinguished professors Cochrane A [1], Guyatt GH [2], and Chalmers IG [3] have led us to the evidence-based medicine (EBM) era. In this era, how to search, critically appraise, and use the best evidence is important. Moreover, systematic review and meta-analysis is the most used tool for summarizing primary data scientifically [4–6] and also the basic for developing clinical practice guideline according to the Institute of Medicine (IOM) [7]. Hence, to perform a systematic review and/ or meta-analysis, assessing the methodological quality of based primary studies is important; naturally, it would be key to assess its own methodological quality before usage. Quality includes internal and external validity, while methodological quality usually refers to internal validity [8, 9]. Internal validity is also recommended as "risk of bias (RoB)" by the Cochrane Collaboration [9].

There are three types of tools: scales, checklists, and items [10, 11]. In 2015, Zeng et al. [11] investigated methodological quality tools for randomized controlled trial (RCT), non-randomized clinical intervention study, cohort study, case-control study, cross-sectional study, case series, diagnostic accuracy study which also called "diagnostic test accuracy (DTA)", animal study, systematic review and meta-analysis, and clinical practice

* Correspondence: zengxiantao1128@163.com; zengxiantao@whucebtm.com
[1]Center for Evidence-Based and Translational Medicine, Zhongnan Hospital, Wuhan University, 169 Donghu Road, Wuchang District, Wuhan 430071, Hubei, China
[2]Department of Evidence-Based Medicine and Clinical Epidemiology, The Second Clinical College, Wuhan University, Wuhan 430071, China
Full list of author information is available at the end of the article

guideline (CPG). From then on, some changes might generate in pre-existing tools, and new tools might also emerge; moreover, the research method has also been developed in recent years. Hence, it is necessary to systematically investigate commonly-used tools for assessing methodological quality, especially those for economic evaluation, clinical prediction rule/model, and qualitative study. Therefore, this narrative review presented related methodological quality (including "RoB") assessment tools for primary and secondary medical studies up to December 2019, and Table 1 presents their basic characterizes. We hope this review can help the producers, users, and researchers of evidence.

Tools for intervention studies

Randomized controlled trial (individual or cluster)

The first RCT was designed by Hill BA (1897–1991) and became the "gold standard" for experimental study design [12, 13] up to now. Nowadays, the Cochrane risk of bias tool for randomized trials (which was introduced in 2008 and edited on March 20, 2011) is the most commonly recommended tool for RCT [9, 14], which is called "RoB". On August 22, 2019 (which was introduced in 2016), the revised revision for this tool to assess RoB in randomized trials (RoB 2.0) was published [15]. The RoB 2.0 tool is suitable for individually-randomized, parallel-group, and cluster- randomized trials, which can be found in the dedicated website https://www.riskofbias.info/welcome/rob-2-0-tool. The RoB 2.0 tool consists of five bias domains and shows major changes when compared to the original Cochrane RoB tool (Table S1A-B presents major items of both versions).

The Physiotherapy Evidence Database (PEDro) scale is a specialized methodological assessment tool for RCT in physiotherapy [16, 17] and can be found in http://www.pedro.org.au/english/downloads/pedro-scale/, covering 11 items (Table S1C). The Effective Practice and Organisation of Care (EPOC) Group is a Cochrane Review Group who also developed a tool (called as "EPOC RoB Tool") for complex interventions randomized trials. This tool has 9 items (Table S1D) and can be found in https://epoc.cochrane.org/resources/epoc-resources-review-authors. The Critical Appraisal Skills Programme (CASP) is a part of the Oxford Centre for Triple Value Healthcare Ltd. (3 V) portfolio, which provides resources and learning and development opportunities to support the development of critical appraisal skills in the UK (http://www.casp-uk.net/) [18–20]. The CASP checklist for RCT consists of three sections involving 11 items (Table S1E). The National Institutes of Health (NIH) also develops quality assessment tools for controlled intervention study (Table S1F) to assess methodological quality of RCT (https://www.nhlbi.nih.gov/health-topics/study-quality-assessment-tools).

The Joanna Briggs Institute (JBI) is an independent, international, not-for-profit researching and development organization based in the Faculty of Health and Medical Sciences at the University of Adelaide, South Australia (https://joannabriggs.org/). Hence, it also develops many critical appraisal checklists involving the feasibility, appropriateness, meaningfulness and effectiveness of healthcare interventions. Table S1G presents the JBI Critical appraisal checklist for RCT, which includes 13 items.

The Scottish Intercollegiate Guidelines Network (SIGN) was established in 1993 (https://www.sign.ac.uk/). Its objective is to improve the quality of health care for patients in Scotland via reducing variations in practices and outcomes, through developing and disseminating national clinical guidelines containing recommendations for effective practice based on current evidence. Hence, it also develops many critical appraisal checklists for assessing methodological quality of different study types, including RCT (Table S1H).

In addition, the Jadad Scale [21], Modified Jadad Scale [22, 23], Delphi List [24], Chalmers Scale [25], National Institute for Clinical Excellence (NICE) methodology checklist [11], Downs & Black checklist [26], and other tools summarized by West et al. in 2002 [27] are not commonly used or recommended nowadays.

Animal study

Before starting clinical trials, the safety and effectiveness of new drugs are usually tested in animal models [28], so animal study is considered as preclinical research, possessing important significance [29, 30]. Likewise, the methodological quality of animal study also needs to be assessed [30]. In 1999, the initial "Stroke Therapy Academic Industry Roundtable (STAIR)" recommended their criteria for assessing the quality of stroke animal studies [31] and this tool is also called "STAIR". In 2009, the STAIR Group updated their criteria and developed "Recommendations for Ensuring Good Scientific Inquiry" [32]. Besides, Macleod et al. [33] proposed a 10-point tool based on STAIR to assess methodological quality of animal study in 2004, which is also called "CAMARADES (The Collaborative Approach to Meta-Analysis and Review of Animal Data from Experimental Studies)"; with "S" presenting "Stroke" at that time and now standing for "Studies" (http://www.camarades.info/). In CAMARADES tool, every item could reach a highest score of one point and the total score for this tool could achieve 10 points (Table S1J).

In 2008, the Systematic Review Center for Laboratory animal Experimentation (SYRCLE) was established in Netherlands and this team developed and released an RoB tool for animal intervention studies - SYRCLE's RoB tool in 2014, based on the original Cochrane RoB

Table 1 The basic characteristics of the included methodological quality (risk of bias) assessment tools

No.	Development Organization	Tool's name	Type of study
1	The Cochrane Collaboration	Cochrane RoB tool and RoB 2.0 tool	Randomized controlled trial Diagnostic accuracy study
2	The Physiotherapy Evidence Database (PEDro)	PEDro scale	Randomized controlled trial
3	The Effective Practice and Organisation of Care (EPOC) Group	EPOC RoB tool	Randomized controlled trial Clinical controlled trials Controlled before-and-after study Interrupted time series studies
4	The Critical Appraisal Skills Programme (CASP)	CASP checklist	Randomized controlled trial Cohort study Case-control study Cross-sectional study Diagnostic test study Clinical prediction rule Economic evaluation Qualitative study Systematic review
5	The National Institutes of Health (NIH)	NIH quality assessment tool	Controlled intervention study Cohort study Cross-sectional study Case-control study Before-after (Pre-post) study with no control group Case-series (Interventional) Systematic review and meta-analysis
6	The Joanna Briggs Institute (JBI)	JBI critical appraisal checklist	Randomized controlled trial Non-randomized experimental study Cohort study Case-control study Cross-sectional study Prevalence data Case reports Economic evaluation Qualitative study Text and expert opinion papers Systematic reviews and research syntheses
7	The Scottish Intercollegiate Guidelines Network (SIGN)	SIGN methodology checklist	Randomized controlled trial Cohort study Case-control study Diagnostic study Economic evaluation Systematic reviews and meta-analyses
8	The Stroke Therapy Academic Industry Roundtable (STAIR) Group	CAMARADES tool	Animal study
9	The SYstematic Review Center for Laboratory animal Experimentation (SYRCLE)	SYRCLE's RoB tool	Animal study
10	Sterne JAC et al.	ROBINS-I tool	Non-randomised interventional study
11	Slim K et al.	MINORS tool	Non-randomised interventional study
12	The Canada Institute of Health Economics (IHE)	IHE quality appraisal tool	Case-series (Interventional)
13	Wells GA et al.	Newcastle-Ottawa Scale (NOS)	Cohort study Case-control study
14	Downes MJ et al.	AXIS tool	Cross-sectional study
15	The Agency for Healthcare Research and Quality (AHRQ)	AHRQ methodology checklist	Cross-sectional/ Prevalence study
16	Crombie I	Crombie's items	Cross-sectional study
17	The Good Research for Comparative Effectiveness (GRACE) Initiative	GRACE checklist	Comparative effectiveness research
18	Whiting PF et al.	QUADAS tool and QUADAS-2 tool	Diagnostic accuracy study

Table 1 The basic characteristics of the included methodological quality (risk of bias) assessment tools *(Continued)*

No.	Development Organization	Tool's name	Type of study
19	The National Institute for Clinical Excellence (NICE)	NICE methodology checklist	Economic evaluation
20	The Cabinet Office	The Quality Framework: Cabinet Office checklist	Qualitative study (social research)
21	Hayden JA et al.	QIPS tool	Prediction study (predictor finding study)
22	Wolff RF et al.	PROBAST	Prediction study (prediction model study)
23	The (COnsensus-based Standards for the selection of health Measurement INstruments) initiative	COSMIN RoB checklist	Patient-reported outcome measure development Content validity Structural validity Internal consistency Cross-cultural validity/ measurement invariance Reliability Measurement error Criterion validity Hypotheses testing for construct validity Responsiveness
24	Shea BJ et al.	AMSTAR and AMSTAR 2	Systematic review
25	The Decision Support Unit (DSU)	DSU network meta-analysis (NMA) methodology checklist	Network meta-analysis
26	Whiting P et al.	ROBIS tool	Systematic review
27	Brouwers MC et al.	AGREE instrument and AGREE II instrument	Clinical practice guideline

AMSTAR A measurement tool to assess systematic reviews, *AHRQ* Agency for healthcare research and quality, *AXIS* Appraisal tool for cross-sectional studies, *CASP* Critical appraisal skills programme, *CAMARADES* The collaborative approach to meta-analysis and review of animal data from experimental studies, *COSMIN* Consensus-based standards for the selection of health measurement instruments, *DSU* Decision support unit, *EPOC* the effective practice and organisation of care group, *GRACE* The god research for comparative effectiveness initiative, *IHE* Canada institute of health economics, *JBI* Joanna Briggs Institute, *MINORS* Methodological index for non-randomized studies, *NOS* Newcastle-Ottawa scale, *NMA* network meta-analysis, *NIH* national institutes of health, *NICE* National institute for clinical excellence, *PEDro* physiotherapy evidence database, *PROBAST* The prediction model risk of bias assessment tool, *QUADAS* Quality assessment of diagnostic accuracy studies, *QIPS* Quality in prognosis studies, *RoB* Risk of bias, *ROBINS-I* Risk of bias in non-randomised studies - of interventions, *ROBIS* Risk of bias in systematic review, *SYRCLE* Systematic review center for laboratory animal experimentation, *STAIR* Stroke therapy academic industry roundtable, *SIGN* The Scottish intercollegiate guidelines network

Tool [34]. This new tool contained 10 items which had become the most recommended tool for assessing the methodological quality of animal intervention studies (Table S1I).

Non-randomised studies

In clinical research, RCT is not always feasible [35]; therefore, non-randomized design remains considerable. In non-randomised study (also called quasi-experimental studies), investigators control the allocation of participants into groups, but do not attempt to adopt randomized operation [36], including follow-up study. According to with or without comparison, non-randomized clinical intervention study can be divided into comparative and non-comparative sub-types, the Risk Of Bias In Non-randomised Studies - of Interventions (ROBINS-I) tool [37] is the preferentially recommended tool. This tool is developed to evaluate risk of bias in estimating comparative effectiveness (harm or benefit) of interventions in studies not adopting randomization in allocating units (individuals or clusters of individuals) into comparison groups. Besides, the JBI critical appraisal checklist for quasi-experimental studies (non-randomized experimental studies) is also suitable, which includes 9 items. Moreover, the methodological index for non-randomized studies (MINORS) [38] tool can also be used, which contains a total of 12 methodological points; the first 8 items could be applied for both non-comparative and comparative studies, while the last 4 items appropriate for studies with two or more groups. Every item is scored from 0 to 2, and the total scores over 16 or 24 give an overall quality score. Table S1K-L-M presented the major items of these three tools.

Non-randomized study with a separate control group could also be called clinical controlled trial or controlled before-and-after study. For this design type, the EPOC RoB tool is suitable (see Table S1D). When using this tool, the "random sequence generation" and "allocation concealment" should be scored as "High risk", while grading for other items could be the same as that for randomized trial.

Non-randomized study without a separate control group could be a before-after (Pre-Post) study, a case series (uncontrolled longitudinal study), or an interrupted time series study. A case series is described a series of individuals, who usually receive the same intervention, and contains non control group [9]. There are several tools for assessing the methodological quality of case series study. The latest one was developed by Moga C et al. [39] in 2012 using a modified Delphi technique, which was developed by the Canada Institute of Health Economics (IHE); hence, it is also called "IHE Quality Appraisal Tool" (Table S1N). Moreover, NIH also develops a quality assessment tool for case series study, including 9 items (Table S1O). For interrupted time series studies, the "EPOC RoB tool for interrupted time series studies" is recommended (Table S1P). For the before-after study, we recommend the NIH quality assessment tool for before-after (Pre-Post) study without control group (Table S1Q).

In addition, for non-randomized intervention study, the Reisch tool (Check List for Assessing Therapeutic Studies) [11, 40], Downs & Black checklist [26], and other tools summarized by Deeks et al. [36] are not commonly used or recommended nowadays.

Tools for observational studies and diagnostic study

Observational studies include cohort study, case-control study, cross-sectional study, case series, case reports, and comparative effectiveness research [41], and can be divided into analytical and descriptive studies [42].

Cohort study

Cohort study includes prospective cohort study, retrospective cohort study, and ambidirectional cohort study [43]. There are some tools for assessing the quality of cohort study, such as the CASP cohort study checklist (Table S2A), SIGN critical appraisal checklists for cohort study (Table S2B), NIH quality assessment tool for observational cohort and cross-sectional studies (Table S2C), Newcastle-Ottawa Scale (NOS; Table S2D) for cohort study, and JBI critical appraisal checklist for cohort study (Table S2E). However, the Downs & Black checklist [26] and the NICE methodology checklist for cohort study [11] are not commonly used or recommended nowadays.

The NOS [44, 45] came from an ongoing collaboration between the Universities of Newcastle, Australia and Ottawa, Canada. Among all above mentioned tools, the NOS is the most commonly used tool nowadays which also allows to be modified based on a special subject.

Case-control study

Case-control study selects participants based on the presence of a specific disease or condition, and seeks earlier exposures that may lead to the disease or outcome [42]. It has an advantage over cohort study, that is the issue of "drop out" or "loss in follow up" of participants as seen in cohort study would not arise in such study. Nowadays, there are some acceptable tools for assessing the methodological quality of case-control study, including CASP case-control study checklist (Table S2F), SIGN critical appraisal checklists for case-control study (Table S2G), NIH quality assessment tool of case-control study (Table S2H), JBI critical appraisal checklist for case-control study (Table S2I), and the NOS for case-control study (Table S2J). Among them,

the NOS for case-control study is also the most frequently used tool nowadays and allows to be modified by users.

In addition, the Downs & Black checklist [26] and the NICE methodology checklist for case-control study [11] are also not commonly used or recommended nowadays.

Cross-sectional study (analytical or descriptive)

Cross-sectional study is used to provide a snapshot of a disease and other variables in a defined population at a time point. It can be divided into analytical and purely descriptive types. Descriptive cross-sectional study merely describes the number of cases or events in a particular population at a time point or during a period of time; whereas analytic cross-sectional study can be used to infer relationships between a disease and other variables [46].

For assessing the quality of analytical cross-sectional study, the NIH quality assessment tool for observational cohort and cross-sectional studies (Table S2C), JBI critical appraisal checklist for analytical cross-sectional study (Table S2K), and the Appraisal tool for Cross-Sectional Studies (AXIS tool; Table S2L) [47] are recommended tools. The AXIS tool is a critical appraisal tool that addresses study design and reporting quality as well as the risk of bias in cross-sectional study, which was developed in 2016 and contains 20 items. Among these three tools, the JBI checklist is the most preferred one.

Purely descriptive cross-sectional study is usually used to measure disease prevalence and incidence. Hence, the critical appraisal tool for analytic cross-sectional study is not proper for the assessment. Only few quality assessment tools are suitable for descriptive cross-sectional study, like the JBI critical appraisal checklist for studies reporting prevalence data [48] (Table S2M), Agency for Healthcare Research and Quality (AHRQ) methodology checklist for assessing the quality of cross-sectional/prevalence study (Table S2N), and Crombie's items for assessing the quality of cross-sectional study [49] (Table S2O). Among them, the JBI tool is the newest.

Case series and case reports

Unlike above mentioned interventional case series, case reports and case series are used to report novel occurrences of a disease or a unique finding [50]. Hence, they belong to descriptive studies. There is only one tool – the JBI critical appraisal checklist for case reports (Table S2P).

Comparative effectiveness research

Comparative effectiveness research (CER) compares real-world outcomes [51] resulting from alternative treatment options that are available for a given medical condition. Its key elements include the study of effectiveness (effect in the real world), rather than efficacy (ideal effect), and the comparisons among alternative strategies [52]. In 2010, the Good Research for Comparative Effectiveness (GRACE) Initiative was established and developed principles to help healthcare providers, researchers, journal readers, and editors evaluate inherent quality for observational research studies of comparative effectiveness [41]. And in 2016, a validated assessment tool – the GRACE Checklist v5.0 (Table S2Q) was released for assessing the quality of CER.

Diagnostic study

Diagnostic tests, also called "Diagnostic Test Accuracy (DTA)", are used by clinicians to identify whether a condition exists in a patient or not, so as to develop an appropriate treatment plan [53]. DTA has several unique features in terms of its design which differ from standard intervention and observational evaluations. In 2003, Penny et al. [53, 54] developed a tool for assessing the quality of DTA, namely Quality Assessment of Diagnostic Accuracy Studies (QUADAS) tool. In 2011, a revised "QUADAS-2" tool (Table S2R) was launched [55, 56]. Besides, the CASP diagnostic checklist (Table S2S), SIGN critical appraisal checklists for diagnostic study (Table S2T), JBI critical appraisal checklist for diagnostic test accuracy studies (Table S2U), and the Cochrane risk of bias assessing tool for diagnostic test accuracy (Table S2V) are also common useful tools in this field.

Of them, the Cochrane risk of bias tool (https://methods.cochrane.org/sdt/) is based on the QUADAS tool, and the SIGN and JBI tools are based on the QUADAS-2 tool. Of course, the QUADAS-2 tool is the first recommended tool. Other relevant tools reviewed by Whiting et al. [53] in 2004 are not used nowadays.

Tools for other primary medical studies

Health economic evaluation

Health economic evaluation research comparatively analyses alternative interventions with regard to their resource uses, costs and health effects [57]. It focuses on identifying, measuring, valuing and comparing resource use, costs and benefit/effect consequences for two or more alternative intervention options [58]. Nowadays, health economic study is increasingly popular. Of course, its methodological quality also needs to be assessed before its initiation. The first tool for such assessment was developed by Drummond and Jefferson in 1996 [59], and then many tools have been developed based on the Drummond's items or its revision [60], such as the SIGN critical appraisal checklists for economic evaluations (Table S3A), CASP economic evaluation checklist (Table S3B), and the JBI critical appraisal checklist for economic evaluations (Table S3C). The

NICE only retains one methodology checklist for economic evaluation (Table S3D).

However, we regard the Consolidated Health Economic Evaluation Reporting Standards (CHEERS) statement [61] as a reporting tool rather than a methodological quality assessment tool, so we do not recommend it to assess the methodological quality of health economic evaluation.

Qualitative study

In healthcare, qualitative research aims to understand and interpret individual experiences, behaviours, interactions, and social contexts, so as to explain interested phenomena, such as the attitudes, beliefs, and perspectives of patients and clinicians; the interpersonal nature of caregiver and patient relationships; illness experience; and the impact of human sufferings [62]. Compared with quantitative studies, assessment tools for qualitative studies are fewer. Nowadays, the CASP qualitative research checklist (Table S3E) is the most frequently recommended tool for this issue. Besides, the JBI critical appraisal checklist for qualitative research [63, 64] (Table S3F) and the Quality Framework: Cabinet Office checklist for social research [65] (Table S3G) are also suitable.

Prediction studies

Clinical prediction study includes predictor finding (prognostic factor) studies, prediction model studies (development, validation, and extending or updating), and prediction model impact studies [66]. For predictor finding study, the Quality In Prognosis Studies (QIPS) tool [67] can be used for assessing its methodological quality (Table S3H). For prediction model impact studies, if it uses a randomized comparative design, tools for RCT can be used, especially the RoB 2.0 tool; if it uses a non-randomized comparative design, tools for non-randomized studies can be used, especially the ROBINS-I tool. For diagnostic and prognostic prediction model studies, the Prediction model Risk Of Bias Assessment Tool (PROBAST; Table S3I) [68] and CASP clinical prediction rule checklist (Table S3J) are suitable.

Text and expert opinion papers

Text and expert opinion-based evidence (also called "non-research evidence") comes from expert opinions, consensus, current discourse, comments, and assumptions or assertions that appear in various journals, magazines, monographs and reports [69–71]. Nowadays, only the JBI has a critical appraisal checklist for the assessment of text and expert opinion papers (Table S3K).

Outcome measurement instruments

An outcome measurement instrument is a "device" used to collect a measurement. The range embraced by the term 'instrument' is broad, and can refer to questionnaire (e.g. patient-reported outcome such as quality of life), observation (e.g. the result of a clinical examination), scale (e.g. a visual analogue scale), laboratory test (e.g. blood test) and images (e.g. ultrasound or other medical imaging) [72, 73]. Measurements can be subjective or objective, and either unidimensional (e.g. attitude) or multidimensional. Nowadays, only one tool - the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) Risk of Bias checklist [74–76] (www.cosmin.nl/) is proper for assessing the methodological quality of outcome measurement instrument, and Table S3L presents its major items, including patient - reported outcome measure (PROM) development (Table S3LA), content validity (Table S3LB), structural validity (Table S3LC), internal consistency (Table S3LD), cross-cultural validity/ measurement invariance (Table S3LE), reliability (Table S3LF), measurement error (Table S3LG), criterion validity (Table S3LH), hypotheses testing for construct validity (Table S3LI), and responsiveness (Table S3LJ).

Tools for secondary medical studies

Systematic review and meta-analysis

Systematic review and meta-analysis are popular methods to keep up with current medical literature [4–6]. Their ultimate purposes and values lie in promoting healthcare [6, 77, 78]. Meta-analysis is a statistical process of combining results from several studies, commonly a part of a systematic review [11]. Of course, critical appraisal would be necessary before using systematic review and meta-analysis.

In 1988, Sacks et al. developed the first tool for assessing the quality of meta-analysis on RCTs - the Sack's Quality Assessment Checklist (SQAC) [79]; And then in 1991, Oxman and Guyatt developed another tool – the Overview Quality Assessment Questionnaire (OQAQ) [80, 81]. To overcome the shortcomings of these two tools, in 2007 the A Measurement Tool to Assess Systematic Reviews (AMSTAR) was developed based on them [82] (http://www.amstar.ca/). However, this original AMSTAR instrument did not include an assessment on the risk of bias for non-randomised studies, and the expert group thought revisions should address all aspects of the conduct of a systematic review. Hence, the new instrument for randomised or non-randomised studies on healthcare interventions - AMSTAR 2 was released in 2017 [83], and Table S4A presents its major items.

Besides, the CASP systematic review checklist (Table S4B), SIGN critical appraisal checklists for systematic reviews and meta-analyses (Table S4C), JBI critical appraisal checklist for systematic reviews and research syntheses (Table S4D), NIH quality assessment tool for

systematic reviews and meta-analyses (Table S4E), The Decision Support Unit (DSU) network meta-analysis (NMA) methodology checklist (Table S4F), and the Risk of Bias in Systematic Review (ROBIS) [84] tool (Table S4G) are all suitable. Among them, the AMSTAR 2 is the most commonly used and the ROIBS is the most frequently recommended.

Among those tools, the AMSTAR 2 is suitable for assessing systematic review and meta-analysis based on randomised or non-randomised interventional studies, the DSU NMA methodology checklist for network meta-analysis, while the ROBIS for meta-analysis based on interventional, diagnostic test accuracy, clinical prediction, and prognostic studies.

Clinical practice guidelines

Clinical practice guideline (CPG) is integrated well into the thinking of practicing clinicians and professional clinical organizations [85–87]; and also make scientific evidence incorporated into clinical practice [88]. However, not all CPGs are evidence-based [89, 90] and their qualities are uneven [91–93]. Until now there were more than 20 appraisal tools have been developed [94]. Among them, the Appraisal of Guidelines for Research and Evaluation (AGREE) instrument has the greatest potential in serving as a basis to develop an appraisal tool for clinical pathways [94]. The AGREE instrument was first released in 2003 [95] and updated to AGREE II instrument in 2009 [96] (www.agreetrust.org/). Now the AGREE II instrument is the most recommended tool for CPG (Table S4H).

Besides, based on the AGREE II, the AGREE Global Rating Scale (AGREE GRS) Instrument [97] was developed as a short item tool to evaluate the quality and reporting of CPGs.

Discussion and conclusions

Currently, the EBM is widely accepted and the major attention of healthcare workers lies in "Going from evidence to recommendations" [98, 99]. Hence, critical appraisal of evidence before using is a key point in this process [100, 101]. In 1987, Mulrow CD [102] pointed out that medical reviews needed routinely use scientific methods to identify, assess, and synthesize information. Hence, perform methodological quality assessment is necessary before using the study. However, although there are more than 20 years have been passed since the first tool emergence, many users remain misunderstand the methodological quality and reporting quality. Of them, someone used the reporting checklist to assess the methodological quality, such as used the Consolidated Standards of Reporting Trials (CONSORT) statement [103] to assess methodological quality of RCT, used the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement [104] to methodological quality of cohort study. This phenomenon indicates more universal education of clinical epidemiology is needed for medical students and professionals.

The methodological quality tool development should according to the characteristics of different study types. In this review, we used "methodological quality", "risk of bias", "critical appraisal", "checklist", "scale", "items", and "assessment tool" to search in the NICE website, SIGN website, Cochrane Library website and JBI website, and on the basis of them, added "systematic review", "meta-analysis", "overview" and "clinical practice guideline" to search in PubMed. Compared with our previous systematic review [11], we found some tools are recommended and remain used, some are used without recommendation, and some are eliminated [10, 29, 30, 36, 53, 94, 105–107]. These tools produce a significant impetus for clinical practice [108, 109].

In addition, compared with our previous systematic review [11], this review stated more tools, especially those developed after 2014, and the latest revisions. Of course, we also adjusted the method of study type classification. Firstly, in 2014, the NICE provided 7 methodology checklists but only retains and updated the checklist for economic evaluation now. Besides, the Cochrane RoB 2.0 tool, AMSTAR 2 tool, CASP checklist, and most of JBI critical appraisal checklists are all the newest revisions; the NIH quality assessment tool, ROBINS-I tool, EPOC RoB tool, AXIS tool, GRACE Checklist, PROBAST, COSMIN Risk of Bias checklist, and ROBIS tool are all newly released tools. Secondly, we also introduced tools for network meta-analysis, outcome measurement instruments, text and expert opinion papers, prediction studies, qualitative study, health economic evaluation, and CER. Thirdly, we classified interventional studies into randomized and non-randomized sub-types, and then further classified non-randomized studies into with and without controlled group. Moreover, we also classified cross-sectional study into analytic and purely descriptive sub-types, and case-series into interventional and observational sub-types. These processing courses were more objective and comprehensive.

Obviously, the number of appropriate tools is the largest for RCT, followed by cohort study; the applicable range of JBI is widest [63, 64], with CASP following closely. However, further efforts remain necessary to develop appraisal tools. For some study types, only one assessment tool is suitable, such as CER, outcome measurement instruments, text and expert opinion papers, case report, and CPG. Besides, there is no proper assessment tool for many study types, such as overview, genetic association study, and cell study. Moreover, existing tools have not been fully accepted. In the future, how to develop well accepted tools remains a significant and important work [11].

Our review can help the professionals of systematic review, meta-analysis, guidelines, and evidence users to choose the best tool when producing or using evidence. Moreover, methodologists can obtain the research topics for developing new tools. Most importantly, we must remember that all assessment tools are subjective, and actual yields of wielding them would be influenced by user's skills and knowledge level. Therefore, users must receive formal training (relevant epidemiological knowledge is necessary), and hold rigorous academic attitude, and at least two independent reviewers should be involved in evaluation and cross-checking to avoid performance bias [110].

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s40779-020-00238-8.

Additional file 1: Table S1. Major components of the tools for assessing intervention studies

Additional file 2: Table S2. Major components of the tools for assessing observational studies and diagnostic study

Additional file 3: Table S3. Major components of the tools for assessing other primary medical studies

Additional file 4: Table S4. Major components of the tools for assessing secondary medical studies

Abbreviations

AGREE GRS: AGREE Global rating scale; AGREE: Appraisal of guidelines for research and evaluation; AHRQ: Agency for healthcare research and quality; AMSTAR: A measurement tool to assess systematic reviews; AXIS: Appraisal tool for cross-sectional studies; CAMARADES: The collaborative approach to meta-analysis and review of animal data from experimental studies; CASP: Critical appraisal skills programme; CER: Comparative effectiveness research; CHEERS: Consolidated health economic evaluation reporting standards; CONSORT: Consolidated standards of reporting trials; COSMIN: Consensus-based standards for the selection of health measurement instruments; CPG: Clinical practice guideline; DSU: Decision support unit; DTA: Diagnostic test accuracy; EBM: Evidence-based medicine; EPOC: The effective practice and organisation of care group; GRACE: The good research for comparative effectiveness initiative; IHE: Canada institute of health economics; IOM: Institute of medicine; JBI: Joanna Briggs Institute; MINORS: Methodological index for non-randomized studies; NICE: National institute for clinical excellence; NIH: National institutes of health; NMA: Network meta-analysis; NOS: Newcastle-Ottawa scale; OQAQ: Overview quality assessment questionnaire; PEDro: Physiotherapy evidence database; PROBAST: The prediction model risk of bias assessment tool; PROM: Patient - reported outcome measure; QIPS: Quality in prognosis studies; QUADAS: Quality assessment of diagnostic accuracy studies; RCT: Randomized controlled trial; RoB: Risk of bias; ROBINS-I: Risk of bias in non-randomised studies - of interventions; ROBIS: Risk of bias in systematic review; SIGN: The Scottish intercollegiate guidelines network; SQAC: Sack's quality assessment checklist; STAIR: Stroke therapy academic industry roundtable; STROBE: Strengthening the reporting of observational studies in epidemiology; SYRCLE: Systematic review center for laboratory animal experimentation

Acknowledgements

The authors thank all the authors and technicians for their hard field work for development methodological quality assessment tools.

Authors' contributions

XTZ is responsible for the design of the study and review of the manuscript; LLM, ZHY, YYW, and DH contributed to the data collection; LLM, YYW, and HW contributed to the preparation of the article. All authors read and approved the final manuscript.

Funding

This work was supported (in part) by the Entrusted Project of National commission on health and health of China (No. [2019]099), the National Key Research and Development Plan of China (2016YFC0106300), and the Nature Science Foundation of Hubei Province (2019FFB03902). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The authors declare that there are no conflicts of interest in this study.

Availability of data and materials

The data and materials used during the current review are all available in this review.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Center for Evidence-Based and Translational Medicine, Zhongnan Hospital, Wuhan University, 169 Donghu Road, Wuchang District, Wuhan 430071, Hubei, China. [2]Department of Evidence-Based Medicine and Clinical Epidemiology, The Second Clinical College, Wuhan University, Wuhan 430071, China. [3]Center for Evidence-Based and Translational Medicine, Wuhan University, Wuhan 430071, China. [4]Global Health Institute, Wuhan University, Wuhan 430072, China.

Received: 17 January 2020 Accepted: 18 February 2020
Published online: 29 February 2020

References

1. Stavrou A, Challoumas D, Dimitrakakis G. Archibald Cochrane (1909-1988): the father of evidence-based medicine. Interact Cardiovasc Thorac Surg. 2013;18(1):121–4.
2. Group E-BMW. Evidence-based medicine. A new approach to teaching the practice of medicine. JAMA. 1992;268(17):2420–5.
3. Levin A. The Cochrane collaboration. Ann Intern Med. 2001;135(4):309–12.
4. Lau J, Ioannidis JP, Schmid CH. Summing up evidence: one answer is not always enough. Lancet. 1998;351(9096):123–7.
5. Clarke M, Chalmers I. Meta-analyses, multivariate analyses, and coping with the play of chance. Lancet. 1998;351(9108):1062–3.
6. Oxman AD, Schunemann HJ, Fretheim A. Improving the use of research evidence in guideline development: 8. Synthesis and presentation of evidence. Health Res Policy Syst. 2006;4:20.
7. Zhang J, Wang Y, Weng H, Wang D, Han F, Huang Q, et al. Management of non-muscle-invasive bladder cancer: quality of clinical practice guidelines and variations in recommendations. BMC Cancer. 2019;19(1):1054.
8. Campbell DT. Factors relevant to the validity of experiments in social settings. Psychol Bull. 1957;54(4):297–312.
9. Higgins J, Green S. Cochrane handbook for systematic reviews of interventions version 5.1.0 [updated March 2011]. The Cochrane Collaboration; 2011.
10. Juni P, Altman DG, Egger M. Systematic reviews in health care: assessing the quality of controlled clinical trials. BMJ. 2001;323(7303):42–6.
11. Zeng X, Zhang Y, Kwong JS, Zhang C, Li S, Sun F, et al. The methodological quality assessment tools for preclinical and clinical studies, systematic review and meta-analysis, and clinical practice guideline: a systematic review. J Evid Based Med. 2015;8(1):2–10.
12. A Medical Research Council Investigation. Treatment of pulmonary tuberculosis with streptomycin and Para-aminosalicylic acid. Br Med J. 1950;2(4688):1073–85.
13. Armitage P. Fisher, Bradford Hill, and randomization. Int J Epidemiol. 2003;32(6):925–8.

14. Higgins JP, Altman DG, Gotzsche PC, Juni P, Moher D, Oxman AD, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ. 2011;343:d5928.
15. Sterne JAC, Savovic J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ. 2019;366: l4898.
16. Maher CG, Sherrington C, Herbert RD, Moseley AM, Elkins M. Reliability of the PEDro scale for rating quality of randomized controlled trials. Phys Ther. 2003;83(8):713–21.
17. Shiwa SR, Costa LO, Costa Lda C, Moseley A, Hespanhol Junior LC, Venancio R, et al. Reproducibility of the Portuguese version of the PEDro scale. Cad Saude Publica. 2011;27(10):2063–8.
18. Ibbotson T, Grimshaw J, Grant A. Evaluation of a programme of workshops for promoting the teaching of critical appraisal skills. Med Educ. 1998;32(5): 486–91.
19. Singh J. Critical appraisal skills programme. J Pharmacol Pharmacother. 2013;4(1):76.
20. Taylor R, Reeves B, Ewings P, Binns S, Keast J, Mears R. A systematic review of the effectiveness of critical appraisal skills training for clinicians. Med Educ. 2000;34(2):120–5.
21. Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? Control Clin Trials. 1996;17(1):1–12.
22. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA. 1995;273(5):408–12.
23. Hartling L, Ospina M, Liang Y, Dryden DM, Hooton N, Krebs Seida J, et al. Risk of bias versus quality assessment of randomised controlled trials: cross sectional study. BMJ. 2009;339:b4012.
24. Verhagen AP, de Vet HC, de Bie RA, Kessels AG, Boers M, Bouter LM, et al. The Delphi list: a criteria list for quality assessment of randomized clinical trials for conducting systematic reviews developed by Delphi consensus. J Clin Epidemiol. 1998;51(12):1235–41.
25. Chalmers TC, Smith H Jr, Blackburn B, Silverman B, Schroeder B, Reitman D, et al. A method for assessing the quality of a randomized control trial. Control Clin Trials. 1981;2(1):31–49.
26. 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 Community Health. 1998; 52(6):377–84.
27. West S, King V, Carey TS, Lohr KN, McKoy N, Sutton SF, et al. Systems to rate the strength of scientific evidence. Evid Rep Technol Assess (Summ). 2002; 47:1–11.
28. Sibbald WJ. An alternative pathway for preclinical research in fluid management. Crit Care. 2000;4(Suppl 2):S8–15.
29. Perel P, Roberts I, Sena E, Wheble P, Briscoe C, Sandercock P, et al. Comparison of treatment effects between animal experiments and clinical trials: systematic review. BMJ. 2007;334(7586):197.
30. Hooijmans CR, Ritskes-Hoitinga M. Progress in using systematic reviews of animal studies to improve translational research. PLoS Med. 2013;10(7): e1001482.
31. Stroke Therapy Academic Industry R. Recommendations for standards regarding preclinical neuroprotective and restorative drug development. Stroke. 1999;30(12):2752–8.
32. Fisher M, Feuerstein G, Howells DW, Hurn PD, Kent TA, Savitz SI, et al. Update of the stroke therapy academic industry roundtable preclinical recommendations. Stroke. 2009;40(6):2244–50.
33. Macleod MR, O'Collins T, Howells DW, Donnan GA. Pooling of animal experimental data reveals influence of study design and publication bias. Stroke. 2004;35(5):1203–8.
34. Hooijmans CR, Rovers MM, de Vries RB, Leenaars M, Ritskes-Hoitinga M, Langendam MW. SYRCLE's risk of bias tool for animal studies. BMC Med Res Methodol. 2014;14:43.
35. McCulloch P, Taylor I, Sasako M, Lovett B, Griffin D. Randomised trials in surgery: problems and possible solutions. BMJ. 2002;324(7351):1448–51.
36. Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakarovitch C, Song F, et al. Evaluating non-randomised intervention studies. Health Technol Assess. 2003;7(27):1–173.
37. Sterne JA, Hernan MA, Reeves BC, Savovic J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. 2016;355:i4919.
38. Slim K, Nini E, Forestier D, Kwiatkowski F, Panis Y, Chipponi J. Methodological index for non-randomized studies (minors): development and validation of a new instrument. ANZ J Surg. 2003;73(9):712–6.
39. Moga C, Guo B, Schopflocher D, Harstall C. Development of a quality appraisal tool for case series studies using a modified delphi technique 2012. http://www.ihe.ca/documents/Case%20series%20studies%20using %20a%20modified%20Delphi%20technique.pdf .(Accept 15 Januray 2020).
40. Reisch JS, Tyson JE, Mize SG. Aid to the evaluation of therapeutic studies. Pediatrics. 1989;84(5):815–27.
41. Dreyer NA, Schneeweiss S, McNeil BJ, Berger ML, Walker AM, Ollendorf DA, et al. GRACE principles: recognizing high-quality observational studies of comparative effectiveness. Am J Manag Care. 2010;16(6):467–71.
42. Grimes DA, Schulz KF. An overview of clinical research: the lay of the land. Lancet. 2002;359(9300):57–61.
43. Grimes DA, Schulz KF. Cohort studies: marching towards outcomes. Lancet. 2002;359(9303):341–5.
44. Wells G, Shea B, O'Connell D, Peterson J, Welch V, Losos M, *et al.* The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. http://www.ohri.ca/programs/clinical_ epidemiology/oxford.asp (Accessed 16 Jan 2020).
45. Stang A. Critical evaluation of the Newcastle-Ottawa scale for the assessment of the quality of nonrandomized studies in meta-analyses. Eur J Epidemiol. 2010;25(9):603–5.
46. Wu L, Li BH, Wang YY, Wang CY, Zi H, Weng H, et al. Periodontal disease and risk of benign prostate hyperplasia: a cross-sectional study. Mil Med Res. 2019;6(1):34.
47. Downes MJ, Brennan ML, Williams HC, Dean RS. Development of a critical appraisal tool to assess the quality of cross-sectional studies (AXIS). BMJ Open. 2016;6(12):e011458.
48. Munn Z, Moola S, Lisy K, Riitano D, Tufanaru C. Methodological guidance for systematic reviews of observational epidemiological studies reporting prevalence and cumulative incidence data. Int J Evid Based Healthc. 2015; 13(3):147–53.
49. Crombie I. Pocket guide to critical appraisal: Oxford. UK: John Wiley & Sons, Ltd; 1996.
50. Gagnier JJ, Kienle G, Altman DG, Moher D, Sox H, Riley D, et al. The CARE guidelines: consensus-based clinical case report guideline development. J Clin Epidemiol. 2014;67(1):46–51.
51. Li BH, Yu ZJ, Wang CY, Zi H, Li XD, Wang XH, et al. A preliminary, multicenter, prospective and real world study on the hemostasis, coagulation, and safety of hemocoagulase bothrops atrox in patients undergoing transurethral bipolar plasmakinetic prostatectomy. Front Pharmacol. 2019;10:1426.
52. Strom BL, Schinnar R, Hennessy S. Comparative effectiveness research. Pharmacoepidemiology. Oxford, UK: John Wiley & Sons, Ltd; 2012. p. 561–79.
53. Whiting P, Rutjes AW, Dinnes J, Reitsma J, Bossuyt PM, Kleijnen J. Development and validation of methods for assessing the quality of diagnostic accuracy studies. Health Technol Assess. 2004;8(25):1–234.
54. 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.
55. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529–36.
56. Schueler S, Schuetz GM, Dewey M. The revised QUADAS-2 tool. Ann Intern Med. 2012;156(4):323.
57. Hoch JS, Dewa CS. An introduction to economic evaluation: what's in a name? Can J Psychiatr. 2005;50(3):159–66.
58. Donaldson C, Vale L, Mugford M. Evidence based health economics: from effectiveness to efficiency in systematic review. UK: Oxford University Press; 2002.
59. Drummond MF, Jefferson TO. Guidelines for authors and peer reviewers of economic submissions to the BMJ. The BMJ economic evaluation working party. BMJ. 1996;313(7052):275–83.
60. Drummond MF, Richardson WS, O'Brien BJ, Levine M, Heyland D. Users' guides to the medical literature. XIII. How to use an article on economic analysis of clinical practice. A. Are the results of the study valid? Evidence-based medicine working group. JAMA. 1997;277(19):1552–7.
61. Husereau D, Drummond M, Petrou S, Carswell C, Moher D, Greenberg D, et al. Consolidated health economic evaluation reporting standards (CHEERS) statement. Value Health. 2013;16(2):e1–5.

62. Wong SS, Wilczynski NL, Haynes RB, Hedges T. Developing optimal search strategies for detecting clinically relevant qualitative studies in MEDLINE. Stud Health Technol Inform. 2004;107(Pt 1):311–6.
63. Vardell E, Malloy M. Joanna briggs institute: an evidence-based practice database. Med Ref Serv Q. 2013;32(4):434–42.
64. Hannes K, Lockwood C. Pragmatism as the philosophical foundation for the Joanna Briggs meta-aggregative approach to qualitative evidence synthesis. J Adv Nurs. 2011;67(7):1632–42.
65. Spencer L, Ritchie J, Lewis J, Dillon L. Quality in qualitative evaluation: a framework for assessing research evidence. UK: Government Chief Social Researcher's office; 2003.
66. Bouwmeester W, Zuithoff NP, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, et al. Reporting and methods in clinical prediction research: a systematic review. PLoS Med. 2012;9(5):1–12.
67. Hayden JA, van der Windt DA, Cartwright JL, Cote P, Bombardier C. Assessing bias in studies of prognostic factors. Ann Intern Med. 2013;158(4):280–6.
68. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med. 2019;170(1):51–8.
69. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ. 1996;312(7023):71–2.
70. Tonelli MR. Integrating evidence into clinical practice: an alternative to evidence-based approaches. J Eval Clin Pract. 2006;12(3):248–56.
71. Woolf SH. Evidence-based medicine and practice guidelines: an overview. Cancer Control. 2000;7(4):362–7.
72. Polit DF. Assessing measurement in health: beyond reliability and validity. Int J Nurs Stud. 2015;52(11):1746–53.
73. Polit DF, Beck CT. Essentials of nursing research: appraising evidence for nursing practice, ninth edition: Lippincott Williams & Wilkins, north American; 2017.
74. Mokkink LB, de Vet HCW, Prinsen CAC, Patrick DL, Alonso J, Bouter LM, et al. COSMIN risk of bias checklist for systematic reviews of patient-reported outcome measures. Qual Life Res. 2018;27(5):1171–9.
75. Mokkink LB, Prinsen CA, Bouter LM, Vet HC, Terwee CB. The consensus-based standards for the selection of health measurement instruments (COSMIN) and how to select an outcome measurement instrument. Braz J Phys Ther. 2016;20(2):105–13.
76. Prinsen CAC, Mokkink LB, Bouter LM, Alonso J, Patrick DL, de Vet HCW, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. Qual Life Res. 2018;27(5):1147–57.
77. Swennen MH, van der Heijden GJ, Boeije HR, van Rheenen N, Verheul FJ, van der Graaf Y, et al. Doctors' perceptions and use of evidence-based medicine: a systematic review and thematic synthesis of qualitative studies. Acad Med. 2013;88(9):1384–96.
78. Gallagher EJ. Systematic reviews: a logical methodological extension of evidence-based medicine. Acad Emerg Med. 1999;6(12):1255–60.
79. Sacks HS, Berrier J, Reitman D, Ancona-Berk VA, Chalmers TC. Meta-analyses of randomized controlled trials. N Engl J Med. 1987;316(8):450–5.
80. Oxman AD. Checklists for review articles. BMJ. 1994;309(6955):648–51.
81. Oxman AD, Guyatt GH. Validation of an index of the quality of review articles. J Clin Epidemiol. 1991;44(11):1271–8.
82. Shea BJ, Grimshaw JM, Wells GA, Boers M, Andersson N, Hamel C, et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. BMC Med Res Methodol. 2007;7:10.
83. Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ. 2017;358:j4008.
84. Whiting P, Savovic J, Higgins JP, Caldwell DM, Reeves BC, Shea B, et al. ROBIS: a new tool to assess risk of bias in systematic reviews was developed. J Clin Epidemiol. 2016;69:225–34.
85. Davis DA, Taylor-Vaisey A. Translating guidelines into practice. A systematic review of theoretic concepts, practical experience and research evidence in the adoption of clinical practice guidelines. CMAJ. 1997;157(4):408–16.
86. Neely JG, Graboyes E, Paniello RC, Sequeira SM, Grindler DJ. Practical guide to understanding the need for clinical practice guidelines. Otolaryngol Head Neck Surg. 2013;149(1):1–7.
87. Browman GP, Levine MN, Mohide EA, Hayward RS, Pritchard KI, Gafni A, et al. The practice guidelines development cycle: a conceptual tool for practice guidelines development and implementation. J Clin Oncol. 1995; 13(2):502–12.
88. Tracy SL. From bench-top to chair-side: how scientific evidence is incorporated into clinical practice. Dent Mater. 2013;30(1):1–15.
89. Chapa D, Hartung MK, Mayberry LJ, Pintz C. Using preappraised evidence sources to guide practice decisions. J Am Assoc Nurse Pract. 2013;25(5):234–43.
90. Eibling D, Fried M, Blitzer A, Postma G. Commentary on the role of expert opinion in developing evidence-based guidelines. Laryngoscope. 2013; 124(2):355–7.
91. Chen YL, Yao L, Xiao XJ, Wang Q, Wang ZH, Liang FX, et al. Quality assessment of clinical guidelines in China: 1993–2010. Chin Med J. 2012; 125(20):3660–4.
92. Hu J, Chen R, Wu S, Tang J, Leng G, Kunnamo I, et al. The quality of clinical practice guidelines in China: a systematic assessment. J Eval Clin Pract. 2013; 19(5):961–7.
93. Henig O, Yahav D, Leibovici L, Paul M. Guidelines for the treatment of pneumonia and urinary tract infections: evaluation of methodological quality using the appraisal of guidelines, research and evaluation ii instrument. Clin Microbiol Infect. 2013;19(12):1106–14.
94. Vlayen J, Aertgeerts B, Hannes K, Sermeus W, Ramaekers D. A systematic review of appraisal tools for clinical practice guidelines: multiple similarities and one common deficit. Int J Qual Health Care. 2005;17(3):235–42.
95. Collaboration A. Development and validation of an international appraisal instrument for assessing the quality of clinical practice guidelines: the AGREE project. Qual Saf Health Care. 2003;12(1):18–23.
96. Brouwers MC, Kho ME, Browman GP, Burgers JS, Cluzeau F, Feder G, et al. AGREE II: advancing guideline development, reporting and evaluation in health care. CMAJ. 2010;182(18):E839–42.
97. Brouwers MC, Kho ME, Browman GP, Burgers JS, Cluzeau F, Feder G, et al. The global rating scale complements the AGREE II in advancing the quality of practice guidelines. J Clin Epidemiol. 2012;65(5):526–34.
98. Guyatt GH, Oxman AD, Kunz R, Falck-Ytter Y, Vist GE, Liberati A, et al. Going from evidence to recommendations. BMJ. 2008;336(7652):1049–51.
99. Andrews J, Guyatt G, Oxman AD, Alderson P, Dahm P, Falck-Ytter Y, et al. GRADE guidelines: 14. Going from evidence to recommendations: the significance and presentation of recommendations. J Clin Epidemiol. 2013; 66(7):719–25.
100. Tunguy-Desmarais GP. Evidence-based medicine should be based on science. S Afr Med J. 2013;103(10):700.
101. Muckart DJ. Evidence-based medicine - are we boiling the frog? S Afr Med J. 2013;103(7):447–8.
102. Mulrow CD. The medical review article: state of the science. Ann Intern Med. 1987;106(3):485–8.
103. Moher D, Schulz KF, Altman D, Group C. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. JAMA. 2001;285(15):1987–91.
104. von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP, et al. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet. 2007;370(9596):1453–7.
105. Sanderson S, Tatt ID, Higgins JP. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. Int J Epidemiol. 2007;36(3):666–76.
106. Willis BH, Quigley M. Uptake of newer methodological developments and the deployment of meta-analysis in diagnostic test research: a systematic review. BMC Med Res Methodol. 2011;11:27.
107. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Group Q-S. A systematic review classifies sources of bias and variation in diagnostic test accuracy studies. J Clin Epidemiol. 2013;66(10):1093–104.
108. Swanson JA, Schmitz D, Chung KC. How to practice evidence-based medicine. Plast Reconstr Surg. 2010;126(1):286–94.
109. Manchikanti L. Evidence-based medicine, systematic reviews, and guidelines in interventional pain management, part I: introduction and general considerations. Pain Physician. 2008;11(2):161–86.
110. Gold C, Erkkila J, Crawford MJ. Shifting effects in randomised controlled trials of complex interventions: a new kind of performance bias? Acta Psychiatr Scand. 2012;126(5):307–14.

ORIGINAL ARTICLE
Iran J Allergy Asthma Immunol
December 2017; 16(6):471-479.

Association Study of CD*226* and CD*247* Genes Single Nucleotide Polymorphisms in Iranian Patients with Systemic Sclerosis

Fatemeh Abbasi[1], Reza Mansouri[1], Farhad Gharibdoost[2], Saeed Aslani[2], Shayan Mostafaei[2], Hoda Kavosi[2], Shiva Poursani[2], Soheila Sobhani[2], and Mahdi Mahmoudi[2]

[1] Immunology Department, Shahid Sadoughi University of Medical Sciences, Yazd, Iran
[2] Rheumatology Research Center, Tehran University of Medical Sciences, Tehran, Iran

Received: 27 May 2017; Received in revised form: 27 July 2017; Accepted: 7 August 2017

ABSTRACT

CD*247* and CD*226* play important roles in signaling of lymphocytes. Single nucleotide polymorphisms (SNPs) of genes encoding CD*247* and CD*226* have been associated with the risk of several autoimmune disorders. This study aimed to evaluate the possible association between *CD226* and *CD247* genes SNPs and risk of systemic sclerosis (SSc) in Iranian population.

Study participants were 455 SSc patients and 455 age, sex, and ethnic -matched healthy individuals. Genotyping of rs2056626 and rs763361 at *CD247* and *CD226* genes, respectively, was carried out using TaqMan MGB-based allelic discrimination real-time PCR. Neither alleles nor genotypes of both SNPs showed significant association with the risk of SSc.

Furthermore, association analysis of the genotypes with clinical manifestations of the disease revealed that rs763361 variants were associated with the forced vital capacity (FVC) in SSc patients.

Our results suggest that genetic variants of *CD226* and *CD247* genes may not be a contributing factor in pathogenesis of SSc in Iranian population.

Keywords: CD*226*; CD*247*; Single nucleotide polymorphism; Systemic sclerosis

INTRODUCTION

Systemic sclerosis (SSc) is a common autoimmune disease characterized by essential vasomotor disturbances, fibrosis, subsequent atrophy of the skin, immunologic abnormalities and autoantibody production.[1,2]

Impaired apoptosis mechanisms of fibroblasts cause prolonged activation of these cells and, therefore, production of cytokines and mediators involved in clinical outcomes of SSc.[3] Many genetic and environmental factors contribute to the pathogenesis of SSc. A huge number of genetic loci has been identified through independent genetic association studies and genome-wide association studies (GWASs) including hundreds of thousands of SNPs located throughout the genome.[4-7]

Corresponding Authors: Mahdi Mahmoudi, Ph.D, and Farhad Gharibdoost, MD;
Rheumatology Research Center, Tehran University of Medical Sciences, Tehran, Iran, PO.Box: 1411713137, Tel/Fax: (+98 21) 8822 0067, E-mails: mahmoudim@tums.ac.ir, gharibdoost@sina.tums.ac.ir

Published by Tehran University of Medical Sciences (http://ijaai.tums.ac.ir)

Although the precise pathogenesis of SSc is still undetermined, it is commonly considered as an autoimmune disorder. During SSc development, autoimmune responses and vasculopathy initiate further events like fibroblast activation and fibrosis.[8,9] As the central culprits of the SSc immunopathogenesis, B cells produce autoantibodies to several autoantigens like those found on endothelial cells. Additionally, ischemia-reperfusion injury due to Raynaud's phenomenon, production of reactive oxygen species (ROS) along with recruitment of inflammatory cell, and subsequent release of inflammatory cytokines induce myofibroblastic transformation and fibroblasts overactivation, eventuating in immoderate collagen synthesis as well as other extracellular compounds.[10,11] As a subset of $CD4^+$ T cells, helper 2 T (Th) cells have been postulated to be involved in aberrant production of profibrotic mediators, like interleukin (IL)-4, IL-13, IL-33, and transforming growth factor β (TGFβ); hence T cells are involved in pathogenesis of SSc through contributing to fibrosis by fibrotic cytokine production, in addition to their role in helping B cells to produce autoantibodies.[8,12,13]

CD*247* (cluster of differentiation *247*), also known as T-cell surface CD3 zeta chain, is part of T cell receptor (TCR)/CD3 complex. The molecule is involved in the assembly as well as transport of the TCR/CD3 complex toward the cell surface. CD*247* plays an essential role in associating the antigen recognition to several intracellular signaling pathways.[14,15] Mutations in *CD247* gene have been demonstrated to be involved in impaired immune function.[16] GWASs demonstrated that an intronic rs2056626 SNP of the *CD247* gene was associated with SSc risk in European and US Caucasians.[4]

CD*226*, known as the DNAX-accessory molecule-1 (DNAM-1), is expressed on immune cells such as natural killer (NK) cells, T cells, NK-T cells, and B cells. The molecule is involved in various biological mechanisms and function of immune cells, particularly as a co-stimulatory molecule in cell signaling.[17,18] *CD226* gene rs763361 SNP is a nonsynonymous variation, which has been frequently occurred in several autoimmune diseases.[19] This polymorphism is attributed to substitution of glycine instead of serine at position 307 (Ser307Gly) in exon-splicing silencer (ESS) region, which may impress expression of CD*226*.[20] The 307Ser variant modifies the splicing of the CD*226* transcript, leading to stimulation of signaling transduction and, therefore, over-activations of T and NK cells.[19]

It appears that genetic variations in *CD226* and *CD247* genes are involved in impaired function of T cells, which are main players in the pathogenesis of SSc. Taking into account the previously reported associations of SNPs in *CD226* and *CD247* genes with a number of autoimmune diseases like systemic sclerosis,[21-24] and that no study has addressed the potential association of polymorphisms in these genes in Iranian population, herein we decided to evaluate the association of rs2056626 and rs763361 at *CD247* and *CD226* genes with the risk of SSc in Iranian population.

MATERIALS AND METHODS

Study Participants

Study population comprised of 455 SSc patients (68 males and 387 females) and 455 healthy controls (67 males and 388 females) with the mean age of 41.55±12.05 and 41.38±12.73, respectively. SSc patients enrolled from the Iranian SSc patients referred to Rheumatology Research Center outpatient clinic, Shariati hospital and diagnosed based on American college of rheumatology (ACR) criteria for SSc. Those patients with past medical history of other autoimmune disorders or family history of SSc were excluded. As the healthy control group, 455 age, sex, and ethnic (Iranian Fars, Turk, Kurd, Lur, and Gilak) -matched individuals were included in this study. Matching the study population ethnically, possibility of spurious results due to population stratification was eliminated.[25] In order to investigate the association between the genotypes of SNPs with SSc phenotypes, the clinical manifestations of the patients were recorded (Table 1). Before sampling, written informed consent was signed by each subject. Ethical Committee of Tehran University of Medical Sciences approved the protocol of the study (No. 93-04-41-27689-290398).

DNA Extraction and Genotyping

Genomic DNA was extracted from 5 mL whole blood samples containing ethylenediaminetetraacetic acid (EDTA) using the standard phenol/chloroform method.[26] The optical density values were used to evaluate the concentration and purity of the extracted DNA (NanoDrop 2000C). All DNA samples were stored at -20°C until further experiments. Study

Table 1. Baseline and clinical data of the studied patients with systemic sclerosis

Characteristic (n=445)	Value
Male/ female	68 (15%)/ 387 (85%)
Age*	41.55±12.05
Disease Duration*	10.5 ± 6.14
Limited SSc/ Diffuse SSc	169 (37%)/ 286 (63%)
Raynaud's phenomenon**	33 (7.25%)
Digital ulcer	345 (75.8%)
Lung fibrosis disease	248 (54.5%)
FVC (Pos>120, Neg<80, Normal range: 80-120)	229 (50.3%)
PAP (Neg<35%, Pos>35%)	77 (17%)
LVEF (Pos<40, Neg:55-70, Borderline:40-55)	55 (12%%)
FANA (Neg>1/100, Pos<1/100, Borderline =1/100)	365 (80%)
ACA (Pos>18, Neg<12, Borderline:12-18 RU/mL)	23 (5%)
ATA (Neg<20 RU/mL , Pos≥20)	317 (69.7%)
ARA	12 (2.6%)
Creatinine*	0.97 ± 0.58
Total protein*	7.55 ± 3.07
ESR*	19.35 ± 18.19

FVC, forced vital capacity; PAP, pulmonary artery pressure; LVEF, left ventricular ejection fraction; FANA, fluorescent anti-nuclear antibodies; ACA, anti-centromere antibodies; ATA, anti-topoisomerase antibody; ARA, Anti-RNA polymerase III; ESR, erythrocyte sedimentation rate.

* Data represented as mean ± SD; ** Positive count reported

Table 2. CD*226* and CD*247* genetic variants analyzed in systemic sclerosis (SSc) patients and healthy controls

SNP	Chromosome	Position	Alleles	Amino acid change
rs763361	18	69864406	C/T	Ser307Gly
rs2056626	1	167451188	G/T	Intron

subjects were genotyped for *CD247* gene rs2056626 and *CD226* gene rs763361 SNPs (Table 2) using the StepOnePlus Real-Time PCR System (Applied Biosystems, Foster City, CA, USA) and allelic discrimination TaqMan MGB-based assays (Applied Biosystems, Foster City, USA). All PCR reactions mixture contained approximately 25-75 ng of DNA, 5 µL Taq-Man Master Mix containing Taq DNA polymerase and dNTPs (Applied Biosystems, Foster City, USA), 0.25 µL Taq-Man Genotyping Assay mix containing primers and FAM or VIC labeled probes (Applied Biosystems, Foster City, USA), and distilled water for a final volume of 10 µL. Thermocyclic conditions of PCR were: initially 60°C for 30 seconds and then 95°C for 10 mins, and subsequently 40 cycles of amplification (95°C for 15 seconds and 60°C for 1 min), and finally 60°C for 30 seconds. Allele calling was performed by analyzing allelic discrimination plots using ABI SDS V 2.3 software (Applied Biosystems, Foster City, USA).

Statistical Analysis

Demographic and clinical characteristics of the study population were assessed by descriptive statistical analysis. The associations between SSc and *CD226* and *CD247* genes SNPs were analyzed by Logistic Regression and $\chi 2$ test or two-tailed Fisher's exact test. Odds ratios (OR) and confidence intervals (95% CI) were employed for risk estimation. *p* values were adjusted by Benjamini-Hochberg Method (BHM) and considered statistically significant if they were less than 0.05. Adherence to the Hardy–Weinberg Equilibrium (HWE) was evaluated using $\chi 2$ test in Package 'genetics' of R-Software (R Core Team, Austria).

RESULTS

Table 1 shows the clinical specifications of the studied SSc population with more details.

Genotype distribution of rs2056626 (p=0.66) and rs763361 (p=0.44) in control subjects did not disclose significant deviation from HWE (Table 3). For both SNPs, the allele with the highest frequency was considered as the reference allele and minor allele frequency (MAF) was reported, according to NCBI database (https://www.ncbi.nlm.nih.gov/snp). The reference genotype was also selected according to the monozygotic genotype with alleles of highest frequency. The C allele of the rs763361 SNP was less represented in SSc patients than controls (39.12% *vs*. 41.75%). However, the frequency difference was not significant (OR= 0.81, CI: 0.55-1.20; p=0.30). The CT genotype of rs763361 SNP had lower frequency in SSc patients compared with healthy controls (47.13% *vs*. 51.15%) and the difference was not significant (OR= 0.80, CI: 0.54-1.21; p=0.30). Alternately, the CC genotype was distributed almost equally between patient and control groups (15.57% *vs*. 16.13%); hence the frequency distribution difference was not statistically significant (OR= 0.84, CI: 0.48-1.47; p=0.55). As the dominant model, the CT+TT genotype had no significant distribution difference between SSc and healthy control groups (62.64% *vs*. 67.25%; OR=0.81, CI: 62-1.07; p=0.15).

For rs2056626 (Table 3), the G allele was found to be highly represented in SSc patients in comparison to controls (38.89% *vs*. 36.87%); however, no significant difference was observed in the allele distribution between patients and controls (OR=1.14, CI: 0.68-1.92; p=0.61). Among the genotypes of rs2056626, both GT and GG genotypes did not show statistically significant differences between study groups. The GG+GT model was assigned as the dominant genotype and its distribution difference was not statistically significant between patients and controls (OR=1.10, CI: 0.84-1.44; p=0.45).

Clinical manifestations of the SSc patients including Raynaud's phenomenon, digital ulcer history, lung fibrosis disease, forced vital capacity (FVC), pulmonary artery pressure (PAP), left ventricular ejection fraction (LVEF), fluorescent anti-nuclear antibody (FANA), anti-centromere antibody (ACA), anti-topoisomerase antibody (ATA), anti-RNA polymerase III (ARA), Creatinine, total protein, erythrocyte sedimentation rate (ESR) were evaluated in relation to genotypes of both SNPs (Table 4). Among them, there was significant correlation of FVC with rs763361 genotypes (p= 0.036).

Table 3. Allele and genotype distribution of *CD226* gene rs763361 SNP and *CD247* gene rs2056626 SNP in the systemic sclerosis patients and healthy controls

SNP	Allele /Genotype	Case (N=455) N (%)	Control (N=455) N (%)	OR (95% CI)	*p*
rs763361	T (Reference)	554 (60.88)	530 (58.24)	-	-
	C	356 (39.12)	380 (41.75)	0.81 (0.55-1.20)	0.30
	TT (Reference)	170 (37.30)	149 (32.72)	-	-
	CT	214 (47.13)	233 (51.15)	0.80 (0.54-1.21)	0.30
	CC	71 (15.57)	73 (16.13)	0.84 (0.48-1.47)	0.55
	CT+TT	285 (62.64)	306 (67.25)	0.81 (0.62-1.07)	0.15
HWE			p= 0.44		
rs2056626	T (Reference)	556 (61.11)	575 (63.13)	-	-
	G	354 (38.89)	335 (36.87)	1.14 (0.68-1.92)	0.61
	TT (Reference)	174 (38.22)	185 (40.55)	-	-
	GT	208 (45.78)	205 (45.16)	1.07 (0.72-1.61)	0.72
	GG	73 (16.00)	65 (14.29)	1.12 (0.68-2.09)	0.54
	GG+GT	281 (61.75)	270 (59.34)	1.10 (0.84-1.44)	0.45
HWE			p= 0.66		

HWE; hardy–weinberg equilibrium

Table 4. Frequencies of rs763361 and rs2056626 genotypes with various clinical features of patients with systemic sclerosis

Clinical Features	rs763361 genotype distribution					rs2056626 genotype distribution				
	Frequency N (%)	CC N (%)	CT N (%)	TT N (%)	*p**	Frequency N (%)	GG N (%)	GT N (%)	TT N (%)	*p**
Raynaud's phenomenon	350 (76.92)	52 (14.86)	158 (45.14)	140 (40.00)	0.313	349 (76.70)	55 (15.76)	159 (45.56)	135 (38.68)	0.339
Digital ulcers	250 (54.95)	39 (15.60)	116 (46.40)	95 (38.00)	0.434	220 (48.35)	37 (16.82)	96 (43.64)	87 (39.54)	0.613
Lung fibrosis disease	179 (39.34)	19 (10.61)	86 (48.04)	74 (41.34)	0.536	162 (35.60)	30 (18.52)	57 (35.18)	75 (46.30)	0.771
FVC	140 (30.77)	19 (13.58)	52 (37.14)	69 (49.28)	**0.036**	258 (56.70)	142 (55.04)	55 (21.32)	61 (23.64)	0.746
PAP	50 (10.98)	5 (10.00)	30 (60.00)	15 (30.00)	0.071	58 (12.74)	8 (13.79)	22 (37.93)	28 (48.28)	0.127
LVEF	35b(77.78)	4 (11.43)	13 (37.14)	18 (51.43)	0.261	26 (5.71)	2 (7.69)	18 (69.23)	6 (23.08)	0.631
FANA	358 (79.56)	52 (14.52)	175 (48.89)	131 (36.59)	0.211	356 (78.24)	59 (16.57)	155 (43.54)	142 (39.89)	0.351
ACA	22 (4.83)	4 (18.18)	13 (59.09)	5 (22.73)	0.869	16 (3.51)	4 (25.00)	8 (50.00)	4 (25.00)	0.108
ATA	310 (68.13)	47 (15.16)	142 (45.81)	121 (39.03)	0.480	295 (64.83)	39 (13.22)	132 (44.75)	124 (42.03)	0.365
ARA	41 (9.01)	7 (17.07)	21 (51.22)	13 (31.71)	0.359	12 (2.64)	2 (16.67)	4 (33.33)	6 (50.00)	0.493
Creatinine	26 (5.71)	2 (7.69)	15 (57.69)	9 (34.62)	0.598	34 (7.47)	10 (29.41)	12 (35.29)	12 (35.29)	0.284
Total protein	47 (10.33)	9 (19.15)	21 (44.68)	17 (36.17)	0.799	46 (10.11)	6 (13.04)	18 (39.13)	22 (47.83)	0.609
ESR	63 (13.85)	11 (17.46)	30 (47.62)	22 (34.92)	0.818	98 (21.54)	16 (16.33)	46 (46.94)	36 (36.73)	0.695

* Benjamini-Hochberg was applied to control the false discovery rate (FDR).

FVC, forced vital capacity; PAP, pulmonary artery pressure; LVEF, left ventricular ejection fraction; FANA, fluorescent anti-nuclear antibodies; ACA, anti-centromere antibodies; ATA, anti-topoisomerase antibody; ARA, Anti-RNA polymerase III; ESR, erythrocyte sedimentation rate

DISCUSSION

SSc is a complex heterogenic disorder of connective tissue and small arteries, defined by the hallmarks of triad of fibrosis, inflammation and vascular injury.[27,28] Fibrosis caused by dermal fibroblast accounts for a wide range of disease outcomes.[29-31] Previous studies classified the genetic variants of SSc in two distinct groups as follows: first, the genetic factors involved in immune system dysfunctions, which many of them have been detected in GWASs. Second, the genetic variants which promote the cellular and molecular mechanisms involved in the progression of inflammation, autoantibody formation and fibrosis development.[32] Based on the pathogenesis of SSc, great attention has been dedicated to the genetics, which affect the immune regulation mechanisms and autoimmunity pathways. Hence, the SNPs in each of these genes might predispose individuals to SSc disease or supply a susceptibility condition for the effect of genetic variants.[33,34]

Accumulating evidence has implicated to a number of immune system perturbations in pathogenesis of SSc. Immune cells, particularly lymphocytes, demonstrate aberrant activation trends in the initial course of SSc pathogenesis. Immune cells, including T cells, macrophages, mast cells, and B cells, infiltrate to skin before any histologic signs of skin fibrosis.[35,36] T cells accumulate in the skin lesions of the SSc patients and demonstrates the signs of activation such as increased expression of IL-2, CD69, and HLA-DR. Additionally, increased serum levels of T-cell associated cytokines like IL-4, IL-13, and IL-17 have been identified in SSc patients.[37] Studies show that both αβ and γδ T cells infiltrate in skin lesions of SSc patients, which is specified only as a consequence of antigen-driven proliferation of T cells.[37] The role of T cells in induction of fibrosis is mediated through production of cytokines or direct cell-cell contact with B cells and fibroblasts. SSc skin is characterized by infiltrating T cells as well as peripheral blood T cells with a predominantly Th2 profile, which mediates the production of profibrotic cytokines like IL-4, IL-13 and TGFβ.[38,39]

Both CD*226* and CD*247* play crucial roles in the stimulation and activation of T cells. It has been found that the CD*247* expression is changed in chronic autoimmune and inflammatory disorders, as decreased

expression of this molecule was associated with impaired immune response.[40-42] Previous investigations revealed the association of CD*247* genetic polymorphisms with susceptibility to systemic lupus erythematosus (SLE), a systemic autoimmune disorder.[43,44] Furthermore, the 3′ untranslated region (3′UTR) of *CD247* gene harbors a number of genetic variations, which have been attributed to downregulation of CD*247*, manifested through immune dysfunction and systemic autoimmunity.[43] CD*247*, which is T-cell surface CD3 zeta molecule, participates in signal transduction in T cells. A GWAS reported that rs2056626 SNP at *CD247* gene is associated with the risk of SSc.[4] Thereafter, this association was validated by an independent cohort study of French Caucasian population and was shown that minor G allele of rs2056626 SNP conferred susceptibility to SSc in dominant model.[45] On the other hand, SSc patients of Han Chinese did not demonstrate significant association between rs2056626 SNP and SSc or its subtypes such as diffuse (dcSSc) and limited (lcSSc) SSc.[46] In compliance with Chinese population, rs2056626 SNP did not impress the SSc risk in Iranian population. Lundholm et al. described that *CD247* might be a prime candidate gene for *Idd28*, a novel susceptible gene for autoimmune disorders such as autoimmune diabetes, by harboring mediators such as CD25 and FoxP3.[47] Therefore, it could be figured out that CD*247* might induce autoimmunity by affecting the cellular environment and could be considered as a predisposing factor rather than a pathogenic variant which might somewhat justify the lack of relation between rs2056626 variant and the risk of SSc in this survey.

CD*226*, a member of immunoglobulin super family, is the main co-stimulator of NK cells, $CD4^+$ and $CD8^+$ T cells, monocytes, platelets and certain B cells, as these cells play notable role in SSc immunopathogenesis of autoimmune disorders.[48,49] While a bulk of studies has demonstrated that the *CD226* gene rs763361 SNP is associated with the risk of several autoimmune diseases, such as rheumatoid arthritis (RA), SSc, and SLE, there are notable reports showing no associations.[19,48,50-52] Evaluation of the polymorphism in Iranian SSc population did not show significant association of this SNP with the risk of the disease. These discrepancies are probably due to inefficient statistical power, small sample sizes, clinical and ethnical heterogeneity. Nonetheless, meta-analysis could be helpful to resolve the inconsistencies and limitations of the disparate individual investigations. For this reason, Song *et al.* performed a meta-analysis in several autoimmune diseases and observed that the *CD226* gene rs763361 SNP reduced the susceptibility to SLE, SSc, and type 1 diabetes (T1D) in Asians, Europeans, and South Americans.[53]

More than 50% of SSc patients all over the world suffer from reduced FVC, which is considered as the main cause of mortality, despite the recent advances in the treatment of SSc.[54,55] In our study, we detected a significant association of FVC with three distinct rs763361 genotypes which may confirm the hypothesis of CD*226* gene variation involvement in reduced FVC of SSc patients.

Our study did not show significant association of CD*226* and CD*247* genetic polymorphisms with the risk of SSc or the presence of clinical manifestations except than FVC. In other words, CD*226* and CD*247* genetic variants may not contribute to SSc pathogenesis in Iranian population. Further studies with large sample size are encouraged in order to assess the contribution between CD*226*, CD*247* and other profibrotic factors in the blood serum of SSc patients. Alternately, considering other genetic variations discerned by GWAS in association with SSc, such as STAT and IRF genes, further studies in Iranian population aiming to dissect the exact role of immune-related molecules underlying SSc etiopathogenesis will be of great interest.

ACKNOWLEDGEMENTS

The work reported herein was supported by a grant from the Deputy of Research, Tehran University of Medical Sciences (Grant No: 93-04-41-27689).

REFERENCES

1. Varga J, Abraham D. Systemic sclerosis: a prototypic multisystem fibrotic disorder. The Journal of clinical investigation. 2007;117(3):557.
2. Hoogen F, Khanna D, Fransen J, Johnson SR, Baron M, Tyndall A, et al. 2013 classification criteria for systemic sclerosis: an American College of Rheumatology/European League against Rheumatism collaborative initiative. Arthritis & Rheumatology. 2013;65(11):2737-47.
3. Jafarinejad-Farsangi S, Farazmand A, Mahmoudi M,

Gharibdoost F, Karimizadeh E, Noorbakhsh F, et al. MicroRNA-29a induces apoptosis via increasing the Bax: Bcl-2 ratio in dermal fibroblasts of patients with systemic sclerosis. Autoimmunity 2015; 48(6):369-78.
4. Radstake TR, Gorlova O, Rueda B, Martin J-E, Alizadeh BZ, Palomino-Morales R, et al. Genome-wide association study of systemic sclerosis identifies CD247 as a new susceptibility locus. Nat Genet 2010; 42(5):426-9.
5. Chairta P, Nicolaou P, Christodoulou K. Genomic and genetic studies of systemic sclerosis: A systematic review. Human Immunology 2017; 78(2):153-65.
6. Aslani S, Mahmoudi M, Karami J, Jamshidi AR, Malekshahi Z, Nicknam MH. Epigenetic alterations underlying autoimmune diseases. Autoimmunity 2016; 49(2):69-83.
7. Abtahi S, Farazmand A, Mahmoudi M, Ashraf-Ganjouei A, Javinani A, Nazari B, et al. IL-1A rs1800587, IL-1B rs1143634 and IL-1R1 rs2234650 polymorphisms in Iranian patients with systemic sclerosis. Int J Immunogenet 2015; 42(6):423-7.
8. Vettori S, Cuomo G, Iudici M, D'Abrosca V, Giacco V, Barra G, et al. Early systemic sclerosis: serum profiling of factors involved in endothelial, T-cell, and fibroblast interplay is marked by elevated interleukin-33 levels. J Clin Immunol 2014; 34(6):663-8.
9. Matucci-Cerinic M, Kahaleh B, Wigley FM. Evidence that systemic sclerosis is a vascular disease. Arthritis Rheum 2013; 65(8):1953-62.
10. Jimenez SA, Piera-Velazquez S. Endothelial to mesenchymal transition (EndoMT) in the pathogenesis of Systemic Sclerosis-associated pulmonary fibrosis and pulmonary arterial hypertension. Myth or reality? Matrix Biol 2016; 51:26-36.
11. Furue M, Mitoma C, Mitoma H, Tsuji G, Chiba T, Nakahara T, et al. Pathogenesis of systemic sclerosis current concept and emerging treatments. Immunol Res 2017; 65(4):1-8.
12. O'reilly S. Role of interleukin-13 in fibrosis, particularly systemic sclerosis. Biofactors 2013; 39(6):593-6.
13. Sakkas LI, Chikanza IC, Platsoucas CD. Mechanisms of disease: the role of immune cells in the pathogenesis of systemic sclerosis. Nat Clin Pract Rheumatol 2006; 2(12):679-85.
14. Irving BA, Chan AC, Weiss A. Functional characterization of a signal transducing motif present in the T cell antigen receptor zeta chain. J Exp Med 1993; 177(4):1093-103.
15. Sussman JJ, Bonifacino JS, Lippincott-Schwartz J, Weissman AM, Saito T, Klausner RD, et al. Failure to synthesize the T cell CD3-ζ chain: structure and function of a partial T cell receptor complex. Cell 1988; 52(1):85-95.
16. Rieux-Laucat F, Hivroz C, Lim A, Mateo V, Pellier I, Selz F, et al. Inherited and somatic CD3 ζ mutations in a patient with T-cell deficiency. N Engl J Med 2006; 354(18):1913-21.
17. Todd JA, Walker NM, Cooper JD, Smyth DJ, Downes K, Plagnol V, et al. Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. Nat Genet 2007; 39(7):857-64.
18. Shibuya K, Shirakawa J, Kameyama T, Honda S-i, Tahara-Hanaoka S, Miyamoto A, et al. CD226 (DNAM-1) is involved in lymphocyte function–associated antigen 1 costimulatory signal for naive T cell differentiation and proliferation. J Exp Med 2003; 198(12):1829-39.
19. Maiti AK, Kim-Howard X, Viswanathan P, Guillén L, Qian X, Rojas-Villarraga A, et al. Non-synonymous variant (Gly307Ser) in CD226 is associated with susceptibility to multiple autoimmune diseases. Rheumatology 2010; 49(7):1239-44.
20. Löfgren SE, Delgado-Vega AM, Gallant CJ, Sánchez E, Frostegård J, Truedsson L, et al. A 3′-untranslated region variant is associated with impaired expression of CD226 in T and natural killer T cells and is associated with susceptibility to systemic lupus erythematosus. Arthritis Rheum 2010; 62(11):3404-14.
21. Broen JC, Coenen MJ, Radstake TR. Genetics of systemic sclerosis: an update. Curr Rheumatol Rep 2012; 14(1):11-21.
22. Jin J, Chou C, Lima M, Zhou D, Zhou X. Systemic sclerosis is a complex disease associated mainly with immune regulatory and inflammatory genes. Open Rheumatol J 2014; 8(1):29-42.
23. Wells AU. Interstitial lung disease in systemic sclerosis. Presse Med 2014; 43(10):e329-e43.
24. Luo Y, Wang Y, Wang Q, Xiao R, Lu Q. Systemic sclerosis: genetics and epigenetics. J Autoimmun 2013; 41:161-7.
25. Balding DJ. A tutorial on statistical methods for population association studies. Nat Rev Genet 2006; 7(10):781.
26. Sambrook J, Fritsch EF, Maniatis T. Molecular cloning. A laboratory Manual. 2nd ed. New York: Cold Spring Harbor Laboratory Press; 1989.
27. Diab S, Dostrovsky N, Hudson M, Tatibouet S, Fritzler MJ, Baron M, et al. Systemic sclerosis sine scleroderma: a multicenter study of 1417 subjects. J Rheumatol 2014; 41(11):2179-85.

28. Distler O, Cozzio A. Systemic sclerosis and localized scleroderma--current concepts and novel targets for therapy. Semin Immunopathol 2016; 38(1):87-95.
29. Jafarinejad-Farsangi S, Farazmand A, Gharibdoost F, Karimizadeh E, Noorbakhsh F, Faridani H, et al. Inhibition of MicroRNA-21 induces apoptosis in dermal fibroblasts of patients with systemic sclerosis. Int J Dermatol 2016; 55(11):1259-67.
30. Karimizadeh E, Gharibdoost F, Motamed N, Jafarinejad-Farsangi S, Jamshidi A, Mahmoudi M. c-Abl silencing reduced the inhibitory effects of TGF-β1 on apoptosis in systemic sclerosis dermal fibroblasts. Mol Cell Biochem 2015; 405(1-2):169-76.
31. Karimizadeh E, Motamed N, Mahmoudi M, Jafarinejad-Farsangi S, Jamshidi A, Faridani H, et al. Attenuation of fibrosis with selective inhibition of c-Abl by siRNA in systemic sclerosis dermal fibroblasts. Arch Dermatol Res 2015; 307(2):135-42.
32. Matucci-Cerinic M, Kahaleh B, Wigley FM. Review: evidence that systemic sclerosis is a vascular disease. Arthritis Rheum 2013; 65(8):1953-62.
33. Fuschiotti P. Current perspectives on the immunopathogenesis of systemic sclerosis. I Immunotargets Ther 2016; 5:21-35.
34. Mahmoudi M, Fallahian F, Sobhani S, Ghoroghi S, Jamshidi A, Poursani S, et al. Analysis of killer cell immunoglobulin-like receptors (KIRs) and their HLA ligand genes polymorphisms in Iranian patients with systemic sclerosis. Clin Rheumatol 2017; 36(4):853-62.
35. Kalogerou A, Gelou E, Mountantonakis S, Settas L, Zafiriou E, Sakkas L. Early T cell activation in the skin from patients with systemic sclerosis. Ann Rheum Dis 2005; 64(8):1233-5.
36. Whitfield ML, Finlay DR, Murray JI, Troyanskaya OG, Chi J-T, Pergamenschikov A, et al. Systemic and cell type-specific gene expression patterns in scleroderma skin. Proc Natl Acad Sci U S A 2003; 100(21):12319-24.
37. Sakkas LI, Platsoucas CD. Is systemic sclerosis an antigen-driven T cell disease? Arthritis Rheum 2004; 50(6):1721-33.
38. Mavalia C, Scaletti C, Romagnani P, Carossino AM, Pignone A, Emmi L, et al. Type 2 helper T-cell predominance and high CD30 expression in systemic sclerosis. Am J Pathol 1997; 151(6):1751.
39. Sakkas LI, Tourtellotte C, Berney S, Myers AR, Platsoucas CD. Increased levels of alternatively spliced interleukin 4 (IL-4δ2) transcripts in peripheral blood mononuclear cells from patients with systemic sclerosis. Clin Diagn Lab Immunol 1999; 6(5):660-4.
40. Krishnan S, Kiang JG, Fisher CU, Nambiar MP, Nguyen HT, Kyttaris VC, et al. Increased caspase-3 expression and activity contribute to reduced CD3ζ expression in systemic lupus erythematosus T cells. J Immunol 2005; 175(5):3417-23.
41. Krishnan S, Warke VG, Nambiar MP, Wong HK, Tsokos GC, Farber DL. Generation and biochemical analysis of human effector CD4 T cells: alterations in tyrosine phosphorylation and loss of CD3ζ expression. Blood 2001; 97(12):3851-9.
42. Krishnan S, Warke VG, Nambiar MP, Tsokos GC, Farber DL. The FcRγ subunit and Syk kinase replace the CD3ζ-chain and ZAP-70 kinase in the TCR signaling complex of human effector CD4 T cells. J Immunol 2003; 170(8):4189-95.
43. Gorman CL, Russell AI, Zhang Z, Graham DC, Cope AP, Vyse TJ. Polymorphisms in the CD3Z gene influence TCRζ expression in systemic lupus erythematosus patients and healthy controls. J Immunol 2008; 180(2):1060-70.
44. Warchoł T, Piotrowski P, Lianeri M, Cieślak D, Wudarski M, Hrycaj P, et al. The CD3Z 844 T> A polymorphism within the 3′-UTR of CD3Z confers increased risk of incidence of systemic lupus erythematosus. Tissue Antigens 2009; 74(1):68-72.
45. Dieudé P, Boileau C, Guedj M, Avouac J, Ruiz B, Hachulla E, et al. Independent replication establishes the CD247 gene as a genetic systemic sclerosis susceptibility factor. Ann Rheum Dis 2011; 70(9):1695-6.
46. Wang J, Yi L, Guo X, He D, Li H, Guo G, et al. Lack of association of the CD247 SNP rs2056626 with systemic sclerosis in Han Chinese. Open Rheumatol J 2014; 8(1):43-5.
47. Lundholm M, Mayans S, Motta V, Lofgren-Burstrom A, Danska J, Holmberg D. Variation in the Cd3 zeta (Cd247) gene correlates with altered T cell activation and is associated with autoimmune diabetes. J Immunol 2010; 184(10):5537-44.
48. Dieude P, Guedj M, Truchetet M-E, Wipff J, Revillod L, Riemekasten G, et al. Association of the CD226 Ser307 variant with systemic sclerosis: Evidence of a contribution of costimulation pathways in systemic sclerosis pathogenesis. Arthritis Rheum 2011; 63(4):1097-105.
49. Bossini-Castillo L, Simeon CP, Beretta L, Broen JC, Vonk MC, Ríos-Fernández R, et al. A multicenter study confirms CD226 gene association with systemic sclerosis-related pulmonary fibrosis. Arthritis Res Ther 2012; 14(2):85.

50. Liu R, Xu N, Wang X, Shen L, Zhao G, Zhang H, et al. Influence of MIF, CD40, and CD226 polymorphisms on risk of rheumatoid arthritis. Mol Biol Rep 2012; 39(6):6915-22.
51. Du Y, Shen LX, Yu LK, Song Y, Zhu JF, Du R. The CD226 gene in susceptibility of rheumatoid arthritis in the Chinese Han population. Rheumatol Int 2012; 32(5):1299-304.
52. Du Y, Tian L, Shen L, Wang F, Yu L, Song Y, et al. Association of the CD226 single nucleotide polymorphism with systemic lupus erythematosus in the Chinese Han population. Tissue antigens 2011; 77(1):65-7.
53. Song G, Bae S, Choi S, Ji J, Lee Y. Association between the CD226 rs763361 polymorphism and susceptibility to autoimmune diseases: a meta-analysis. Lupus 2012; 21(14):1522-30.
54. Michelfelder M, Becker M, Riedlinger A, Siegert E, Dromann D, Yu X, et al. Interstitial lung disease increases mortality in systemic sclerosis patients with pulmonary arterial hypertension without affecting hemodynamics and exercise capacity. Clin Rheumatol 2016; 36(2):381-90.
55. Man A, Davidyock T, Ferguson LT, Ieong M, Zhang Y, Simms RW. Changes in forced vital capacity over time in systemic sclerosis: application of group-based trajectory modelling. Rheumatology (Oxford) 2015; 54(8):1464-71.

Received: 5 July 2017 | Revised: 13 January 2018 | Accepted: 22 January 2018

DOI: 10.1111/jep.12889

ORIGINAL ARTICLE

WILEY Journal of Evaluation in Clinical Practice
International Journal of Public Health Policy and Health Services Research

Critical appraisal of nonrandomized studies—A review of recommended and commonly used tools

Joan M. Quigley MSc, Lead Systematic Reviewer |
Juliette C. Thompson BSc (Hons), Lead Systematic Reviewer |
Nicholas J. Halfpenny MSci, Senior Systematic Reviewer |
David A. Scott MSc, Executive Principal, Global Head

ICON Health Economics and Epidemiology, ICON Health Economics, Abingdon, UK

Correspondence
Joan M. Quigley, Health Research Board, Grattan House, 67-72 Lower Mount Street, Dublin 2, D02 H638, Republic of Ireland.
Email: jquigley@hrb.ie

Present Address
Joan M. Quigley, Health Research Board, Grattan House, 67-72 Lower Mount Street, Dublin 2, D02 H638, Republic of Ireland.

David A. Scott, Honorary Visiting Professor, Diabetes Research Centre, University of Leicester, Leicester General Hospital, Leicester, LE5 4PW, UK.

Abstract

Rationale, aims, and objectives: When randomized controlled trial data are limited or unavailable, or to supplement randomized controlled trial evidence, health technology assessment (HTA) agencies may rely on systematic reviews of nonrandomized studies (NRSs) for evidence of the effectiveness of health care interventions. NRS designs may introduce considerable bias into systematic reviews, and several methodologies by which to evaluate this risk of bias are available. This study aimed to identify tools commonly used to assess bias in NRS and determine those recommended by HTA bodies.

Methods: Appraisal tools used in NRS were identified through a targeted search of systematic reviews (January 2013-March 2017; MEDLINE and EMBASE [OVID SP]). Recommendations for the critical appraisal of NRS by expert review groups and HTA bodies were reviewed.

Results: From the 686 studies included in the narrative synthesis, 48 critical appraisal tools were identified. Commonly used tools included the Newcastle-Ottawa Scale, the methodological index for NRS, and bespoke appraisal tools. Neither the Cochrane Handbook nor the Centre for Reviews and Dissemination recommends a particular instrument for the assessment of risk of bias in NRS, although Cochrane has recently developed their own NRS critical appraisal tool. Among HTA bodies, only the Canadian Agency for Drugs and Technologies in Health recommends use of a specific critical appraisal tool—SIGN 50 (for cohort or case-control studies). Several criteria including reporting, external validity, confounding, and power were examined.

Conclusion: There is no consensus between HTA groups on the preferred appraisal tool. Reviewers should select from a suite of tools on the basis of the design of studies included in their review.

KEYWORDS

evaluation, evidence-based medicine, health economics, health care, medical informatics, systematic reviews

Abbreviations: ACROBAT-NRSI, A Cochrane Risk Of Bias Assessment Tool: for Non-Randomized Studies of Interventions; AMCP, Academy of Managed Care Pharmacy; AWMSG, All Wales Medicines Strategy Group; CADTH, Canadian Agency for Drugs and Technologies in Health; CASP, Critical Appraisal Skills Programme; CRD, Centre for Reviews and Dissemination; EPOC, Effective Practice and Organisation of Care; GRACE, Good ReseArch for Comparative Effectiveness; GRADE, Grades of Recommendation, Assessment, Development, and Evaluation; HAS, Haute Autorité de santé; HTA, Health technology assessment; ISPOR, International Society for Pharmacoeconomics and Outcomes Research; IQWiG, Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen; JBI-MAStARI, Joanna Briggs Institute Meta Analysis of Statistics Assessment and Review Instrument; MINORS, methodological index for nonrandomized studies; NCPE, National Centre for Pharmacoeconomics; NICE, National Institute for Health and Care Excellence; NOS, Newcastle-Ottawa Scale; NRS, nonrandomized studies; PBAC, Pharmaceutical Benefits Advisory Committee; PEDro, Physiotherapy Evidence Database; PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses; RCT, randomized controlled trial; ROB, risk of bias; RoBANS, risk of bias assessment tool for nonrandomized studies; SIGN, Scottish Intercollegiate Guidelines Network; SMC, Scottish Medicines Consortium; STROBE, Strengthening the Reporting of Observational Studies in Epidemiology; TREND, Transparent Reporting of Evaluations with Nonrandomized Designs; USPSTF, United States Preventive Services Task Force grading system.

1 | INTRODUCTION

The changing paradigm for health care decision making has begun to welcome observational and real-world data in addressing evidence gaps.[1,2] This move to observational data has been fuelled by greater availability and access to electronic health care data and long-term observational studies. Where randomized controlled trials (RCTs) are unavailable, unethical, or implausible, and to supplement RCT evidence, health technology assessment (HTA) agencies commonly rely on nonrandomized studies (NRS) to provide evidence of the effectiveness of health care interventions.[3,4]

Interventions for rare diseases are often approved with short-term trials and with rapid review to give patients access to drugs more quickly. However, in these cases, long-term observational data/postmarketing studies are paramount to illustrating the long-term benefits or side effects of drug use and are commonly being mandated in marketing authorization.[5]

Whilst there is no doubt of the benefits of including NRS in systematic reviews, incorporation of this type of evidence into a systematic review or a meta-analysis requires careful consideration as, due to their design, NRS are especially susceptible to confounding and other types of bias. For example, allocation of patients to an intervention in a nonrandomized fashion can result in selection bias with a disproportionate distribution of known prognostic factors between treatment arms,[6] whereas performance bias may result in cases where study investigators are aware of treatment allocation.[7] Further bias may be introduced by misclassification of or deviation from an intervention, missing data, mismeasurement or misclassification of outcomes, or selective reporting. Whilst RCTs are also susceptible to these potential sources of bias, several tools including those from Cochrane,[7] the Centre for Reviews and Dissemination (CRD),[8] and the Scottish Intercollegiate Guidelines Network (SIGN)[9] are commonly used to critically appraise studies included in a systematic review of RCTs. All published Cochrane reviews must perform an analysis with the Cochrane risk of bias (ROB) tool,[7] making it the most recognized of these tools. A recent evaluation found the Cochrane ROB tool has become the preferred approach to assess ROB in RCTs even in non-Cochrane reviews; however, it is not always implemented in the recommended way.[10]

Although tools for the evaluation of NRS have existed for some time, to date, there is no consensus about which is the most suitable for use for a particular study design. For instance, although a prospective nonrandomized interventional study and a retrospective case series are both NRS by definition, there are substantial differences in their study designs and potential sources of bias that may be difficult to assess with a single tool, most of which assess only a limited spectrum of possible sources of bias. However, researchers would benefit from a consensus on a small suite of tools, which could be used for NRS, dependent on design of studies in the review. This would allow for better interpretation of the results of the ROB analysis and limit the variability in bias judgements, which have been identified as a problem in Cochrane reviews.[11]

We undertook a review of the literature with the aims of identifying the most commonly used and accepted critical appraisal tools for NRS and of providing an overview of the most frequent characteristics of such tools.

2 | METHODS

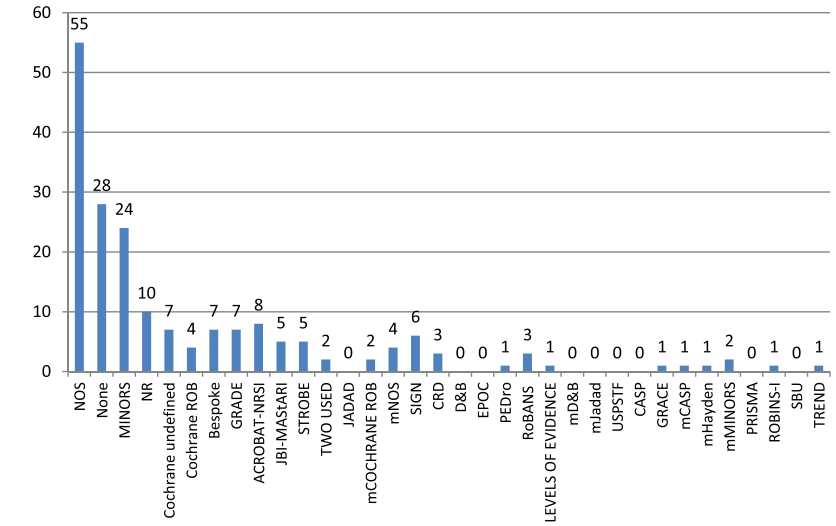
We used a 2-part approach to address our research question. Firstly, we conducted a targeted search in MEDLINE, MEDLINE In-process and Other Non-Indexed Citations, and EMBASE (OVID SP) to identify systematic reviews that included NRS. Abstracts for the records retrieved using the search terms shown in Table S1 were reviewed by a single reviewer. To be eligible for inclusion in the narrative synthesis, reviews were required to be systematic reviews of medical interventions (vs alternative medical interventions or vs no intervention) that included NRS and that were published in peer-reviewed journals in English between January 2013 and March 2017. This pragmatic date range was selected to keep the number of studies returned for review to a manageable number, whilst still providing a representative view of the current state of the literature. A single reviewer confirmed that a systematic review process had been followed in the identified articles and recorded the use of any critical appraisal tool and, where reported, identified which tool was used together with its constituent components.

Secondly, we reviewed recommendations for the critical appraisal of NRS by expert review groups (Cochrane Handbook, Cochrane Bias Methods Group, Cochrane NRS Methods Group, CRD, and SIGN) and HTA bodies (National Institute for Health and Care Excellence [NICE], Scottish Medicines Consortium [SMC], National Centre for Pharmacoeconomics [NCPE], All Wales Medicines Strategy Group [AWMSG], Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen [IQWiG], Pharmaceutical Benefits Advisory Committee [PBAC], Academy of Managed Care Pharmacy [AMCP], Haute Autorité de santé [HAS], and Canadian Agency for Drugs and Technologies in Health [CADTH]). The searches were conducted by visiting the official website of these groups and reviewing the appropriate methodological advice.

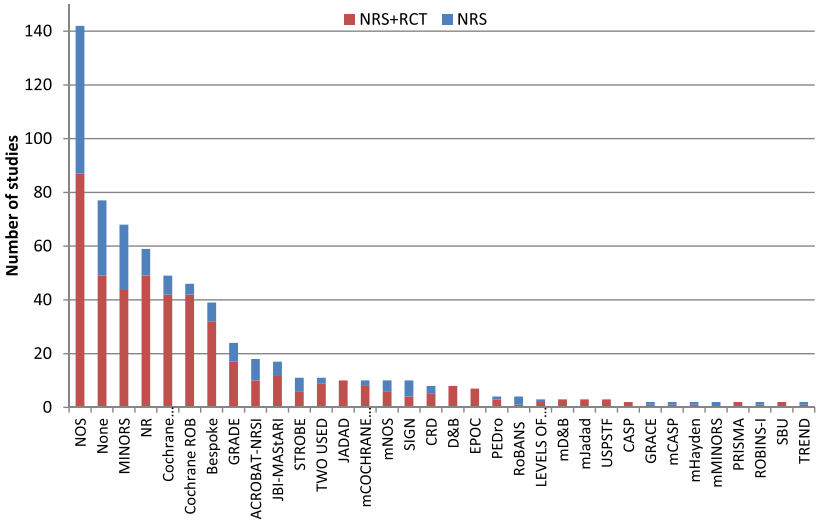
3 | RESULTS

3.1 | Systematic search and use of critical appraisal tools

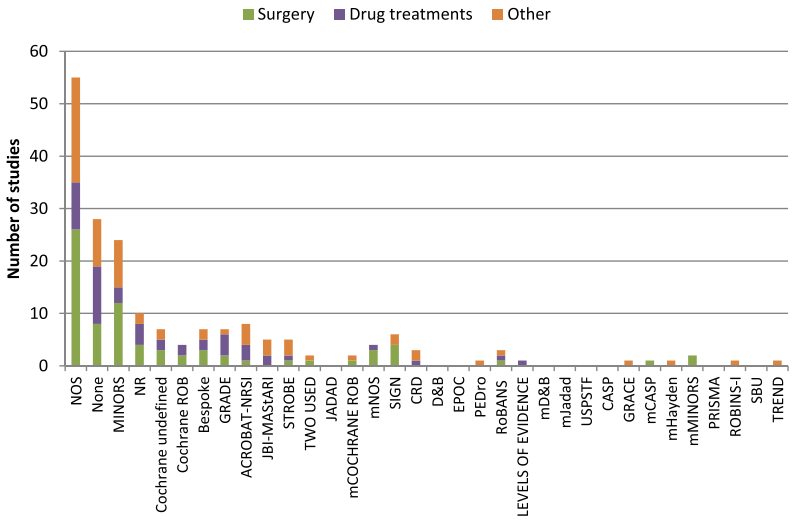
Our search identified 2498 references for screening. Following full-paper review, 686 systematic reviews that included NRS were identified (PRISMA diagram, Figure S1), and a total of 48 different critical appraisal tools were identified. The tools used in more than 2 studies are illustrated in Figure 1. (Remaining identified tools are listed in Appendix S1.) Figure 1A splits the use of each tool by those studies, which considered only NRS (197), and Figure 1B splits to those which included both NRS and RCT (489). A similar pattern of tool usage was observed regardless of study type included. Figure 1C splits studies by intervention type: drug treatments, surgery, and other treatments (eg, medical devices, diagnostic tests, physiotherapy, nutritional support, and mixed interventions). No obvious trends were observed in choice of tool by intervention type. Notably, the methodological index for NRS (MINORS), the surgery-specific tool, was also applied to studies of other interventions. As study design was not a search criterion, most identified reviews included both RCT and NRS



(A) Use of tools split by study designs included in the SR, NRS only



(B) Use of tools split by study designs included in the SR, NRS only or NRS + RCT



(C) Use of tools split by intervention type

FIGURE 1 Frequency of appraisal tool use among 686 systematic reviews that included NRS

evidence, both of which were frequently evaluated using the same critical appraisal tool: The Cochrane ROB tool for RCT[7] was used by 46 of the 686 studies identified by our search. Modified versions of the Cochrane tool were also used although details of these modifications were rarely reported. In addition, Cochrane has multiple published critical appraisal tools (Cochrane ROB, ACROBAT-NRSI, and

ROBINS-I), and often studies did not clarify which tool was used; in these cases, we have recorded the tool as "Cochrane undefined" (49 of 686). The SIGN and CRD tools were used for NRS appraisal in 10 and 8 studies, respectively, although these were also designed for use on trials with a randomized controlled design only.[8,12] For reviews that used an evidence grading tool, eg, GRADE, the review applied the criteria to a within-study comparison, and so they were considered relevant to this research question.

Of the 686 systematic reviews identified by our search strategy, 77 (11%) reported that no appraisal tool was used, and 59 (9%) did not report on the appraisal tool. Among the remaining studies, the most commonly used critical appraisal tool for NRS was the Newcastle-Ottawa Scale for observational studies, reported in 142 of 686 (21%) of the systematic reviews. The MINORS, bespoke tools, GRADE, ACROBAT-NRSI (which has since been developed into ROBINS-I), and JBI-MAStARI tools were also used by multiple studies: 68 (10%), 39 (6%), 24 (3%), 18 (3%), and 17 (2%), respectively. All remaining tools identified were used by less than 2% of the included studies. A full list of the studies included is given in Appendix S1.

Whilst a majority of studies gave a narrative summary of quality assessment, others adjusted for study quality in meta-analysis. Authors typically applied a cut-off in terms of score or domain to classify poorer quality studies and then conducted a sensitivity analysis excluding those studies.

3.2 | Appraisal tools recommended by HTA and professional groups

We reviewed the recommendations for the critical appraisal of NRS from major HTA bodies and institutions involved in developing systematic review methodologies, including Cochrane,[7] the CRD,[8] and the SIGN[12] (Table 1). CADTH was the only HTA group to recommend a specific critical appraisal tool—that produced by the SIGN.[14] Other HTA groups reported less defined recommendations: NICE recommends that an appropriate and validated tool is used,[15] and PBAC requests that relevant documentation is provided to support the use of any tool used.[16] The process of validation that should be undertaken for the critical appraisal tool is not clearly defined. The remaining HTA groups (SMC, NCPE, AWMSG, IQWiG, AMCP, and HAS) did not mention the critical appraisal of NRS. Whilst there is recognition by HTA agencies that NRS can provide valuable information to supplement robust RCT evidence and NRS are increasingly being incorporated in HTA submissions, there is a lack of guidance or consensus as to how this data should be critically appraised.[13,17]

The recognized methodological groups, SIGN[9] and Critical Appraisal Skills Programme,[18] produce different critical appraisal tools for NRS (Table 1), describing several different appraisal tools based on study design. The most recent edition of the Cochrane Handbook (2011) highlights the Downs and Black[19] and Newcastle-Ottawa Scale[20] without recommending either; however, it is noted that of the two, the Newcastle-Ottawa Scale is easier to implement, because of the smaller number of domains.[7] Since the last published edition of the handbook (2011), Cochrane has developed the ROB In NRS–of Interventions (ROBINS-I) tool for the critical appraisal of NRS. The CRD did not make specific recommendations.

3.3 | Overview of components of appraisal tools

An overview of the 5 most commonly used appraisal tools is given in Table 2, which includes only those tools specifically designed for the critical appraisal of NRS. This summary may help review authors to select a critical appraisal tool for use in their systematic review. Two tools that were highly ranked, but not included here, were an undefined Cochrane tool (which could be either the Cochrane ROB tool for RCTs (not capitalised) or that from the NRS group) and the Cochrane ROB tool for RCTs. These are not included as tools to appraise RCTs are unlikely to be appropriate or sufficient for the appraisal of NRS.[21] Questions regarding methods of randomization are not applicable to NRS, and details regarding comparability of the study population may be insufficient.

To aid in interpretation and comparability, we divided the appraisal tool items into 12 domains according to the different questions posed that were common across all tools:

1. Appropriate study design—a clear aim that is precise and relevant in the context of the literature is stated.
2. Patient selection/ inclusion criteria—a clear inclusion criterion is stated.

TABLE 1 HTA and methodological group recommendations

Methodological Groups	Recommendation
CADTH	Recommend a specific critical appraisal tool—which is produced by SIGN[11]
CASP, Oxford, UK	Provide several appraisal tools for NRS including CASP diagnostic checklist, CASP case-control checklist, CASP cohort study checklist, and CASP clinical prediction rule checklist[16]
Cochrane[a]	Produced the ACROBAT-NRSI tool and ROBINS-I tool[9]
Cochrane Handbook (last revised 2011 and so does not capture ongoing work above)	No formal recommendation: highlights Downs and Black as well as Newcastle-Ottawa Scale. States Newcastle-Ottawa easier to implement[7]
CRD	Acknowledge abundance of tools but no formal recommendation as to which to use[9]
SIGN	Provides appraisal tools for various types of NRS including; methodology checklist: cohort studies, methodology checklist case-control studies, and methodology checklist: diagnostic studies[8]

Abbreviations: ACROBAT-NRSI, A Cochrane Risk Of Bias Assessment Tool: for Non-Randomized Studies of Interventions; ADTH, Canadian Agency for Drugs and Technologies in Health; CASP, Critical Appraisal Skills Programme; CRD, Centre for Reviews and Dissemination; HTA, health technology assessment; NRS, nonrandomized studies; ROBINS-I, Risk of Bias in Non-randomized Studies–of Interventions; SIGN, Scottish Intercollegiate Guidelines Network.

[a]An ongoing collaboration between the Cochrane Bias Methods Group and the Cochrane Non-Randomized Studies Methods Group.

TABLE 2 Overview of NRS appraisal tools for use in a systematic review (number of questions relating to each domain)

	Designed to assess	Total items	1. Appropriate design	2. Patient selection/ inclusion criteria	3. Blinding of patients and personnel	4. Assessments of outcomes/ exposure	5. Follow-up/ handling of missing data	6. Reporting (including selective reporting)	7. Subject comparability
ROBINS-I[a]	NRS of interventions	34	…	5	3	10	5	3	…
JBI-MAStARI	RCT and pseudo-RCTs[b]	10	…	2	2	3	…	…	2
	Comparable cohort and case-control studies	9	…	…	…	2	2	…	2
	Descriptive and case series studies	9	…	…	2	2	2	1	…
MINORS	NRS of surgical interventions	12	1	2	…	1	2	…	2
NOS	Case-control studies and cohort studies (2 variations)	8	…	4	…	3	…	…	1
SIGN	Cohort studies	18	1	5	…	6	…	…	…
	Case-control studies	15	1	6	…	2	…	…	…

Abbreviations: ACROBAT-NRSI, A Cochrane Risk Of Bias Assessment Tool: for Non-Randomized Studies of Interventions; JBI-MAStARI, Joanna Briggs Institute Meta-Analysis of Statistics Assessment and Review Instrument; MINORS, methodological index for nonrandomized studies; NOS, Newcastle-Ottawa Scale; NRS, nonrandomized studies; RCT, randomized controlled trial; ROBINS-I, Risk Of Bias in Non-randomized Studies—of Interventions; SIGN, Scottish Intercollegiate Guidelines Network.

[a]We present the ROBINS-I tool as it is an updated version of the ACROBAT-NRSI tool.

[b]Inappropriate randomization technique used.

TABLE 2 (Continued)

	Designed to assess	8. Appropriate end points	9. Confounding	10. Prospective calculation of study size	11. Appropriate statistical analysis	12. Overall study assessment	Scoring scale	Answering method	Validated	Validation technique
ROBINS-I[a]	NRS of interventions		8				Domain-based evaluation	Serious risk; moderate risk; low risk; qualifying statements provided	Unclear	NR
JBI-MAStARI	RCT and pseudo-RCTs[b] Comparable cohort and case-control studies Descriptive and case series studies	... ...	... 2 1	... ...	1 1 1	... ...	Subject to interassessor variability due to criteria being perceived as more or less important. These decisions about the scoring system and any cut-off for inclusion should be made in advance and be agreed upon by all participating reviewers before critical appraisal commences	Met, not met, unclear, inappropriate	Unclear	NR
MINORS	NRS of surgical interventions	1	1	1	1	...	Up to 16 for noncomparative studies up to 24 for comparative studies	Not reported (0), reported but inadequate (1), reported and adequate (2)	Unclear	90 surgeons in France with different specialties
NOS	Case-control studies and cohort studies (2 variations)	...	...	...	...	...	Maximum of one star for each numbered item within the selection and exposure categories maximum of 2 stars can be awarded for comparability category	Several responses provided for each question—must select one	Yes	Content validity and inter-rater reliability established Criterion validity and intra-rater reliability currently being examined
SIGN	Cohort studies Case-control studies	... ...	1 1	... ...	1 1	4 4	One question on the overall methodological quality of the study: High quality (++) Low quality (0)	For the individual items, the answers are yes, no, and unclear	Unclear	"Subjected to detailed evaluation and adaptation to meet SIGN's requirements"

3. Blinding of patients and personnel—details of blinding are provided.
4. Assessment of outcomes/exposure—were outcomes/exposure measured in a reliable way or using an objective criterion?
5. Follow up/handling of missing data—was length of follow-up sufficient? Were patients who withdrew from the study described and included in the analysis?
6. Reporting—are there sufficient descriptions of the groups given to allow comparisons? Any evidence of selective reporting of outcomes?
7. Subject comparability—are the patients included reflective of the total patient population? Are they at a similar point in the course of the disease?
8. Appropriate end points—were the end points appropriate?
9. Confounding—are confounding factors identified and are strategies in place to deal with them?
10. Prospective calculation of study size—how have the sample size power calculations been conducted?
11. Appropriate statistical analysis—has appropriate statistical analysis been conducted?
12. Overall study assessment—is the overall study of adequate quality?

Table 2 shows important domains of critical appraisal of NRS including assessment of patient selection, assessment of outcomes, and whether appropriate statistical analysis has been conducted. These domains are based on descriptions of bias as given in the CRD guidance and the Cochrane Handbook.[7,8] Identification of confounding and methods used to address it were also commonly queried.

The ROBINS-I tool is designed to evaluate the ROB in estimates of the comparative effectiveness of interventions in studies where participants were not randomly allocated.[6] It was developed as a collaboration between the Cochrane Bias Methods Group and the Cochrane NRS for Interventions Methods Group. It is composed of 7 core domains, which then have up to 8 individual points to be assessed; there are a total of 34 questions over the tool's 22 pages. Each domain is rated as low, moderate, serious, or critical ROB with the overall study rating taking the rating of the worst performing domain. The tool is also accompanied by 53 pages of detailed guidance.[22]

The Joanna Briggs Institute (JBI) at the University of Adelaide has 3 versions of the Meta-Analysis of Statistics Assessment and Review Instrument with appropriate usage dependent on the design of included studies. Randomized and quasi-RCTs are considered together with 10 questions for critical appraisal; there are 9 questions to critically appraise comparable cohort/case-control studies and 9 for descriptive/case series studies. The critical appraisal is incorporated into the analytical module of the JBI systematic review software, and guidance is provided for each instrument.[23]

The MINORS was designed to assess the quality of surgical intervention studies. The MINORS is composed of 8 methodological items applicable to all NRS and 4 additional items, which should be applied in the case of comparative studies. Each item scores 2 for reported and adequate, 1 for reported but inadequate, or 0 for not reported, giving a maximum score of 16 for noncomparative studies or 24 for comparative studies.[24]

The Newcastle-Ottawa Quality Assessment Scale was developed as a collaboration between the Universities of Newcastle, Australia, and Ottawa, Canada, and includes separate instruments for case-control and cohort studies. There are 3 domains: selection (4 questions), comparability (1 question), and exposure (3 questions), giving a total of 8 questions. Studies can score up to 9 stars, with up to 2 stars being awarded for comparability. A brief coding manual accompanies each instrument.

The SIGN has a range of methodology checklists; for NRS, one addresses case-control studies, and a second covers cohort studies. For cohort studies, there are 18 statements listed that reviewers check against the study; some statements are only applicable in certain cases, eg, in prospective studies, and these are indicated. For case-control studies, there are 15 statements.[9] Notes are provided for each checklist.

4 | DISCUSSION

This review of critical appraisal tools identified 48 different tools that were used in at least one published review. We have highlighted similarities and differences between the most commonly used tools designed specifically for use with NRS. The complexity and relevance of the tools were highly variable. Some appraisal tools included a simple answering mechanism: "yes" or "no," thus facilitating rapid completion (eg, GRACE, and JBI). Tools for assessing the quality of RCTs were frequently applied inappropriately to NRS. Whilst Cochrane does state that the ROB tool for RCTs may be useful to assess some aspects of bias in controlled studies and prospective cohort studies, this tool may not be appropriate to capture all important aspects of bias or for other NRS designs.[7] The publication of the ROBINS-I tool, and its use in Cochrane systematic reviews, may change the future landscape of commonly used NRS critical appraisal tools and is likely to be favoured in a forthcoming edition of the handbook since the current guidance is based on review from 2003.[25] As a new tool, there is a lack of published critique on its usage. However, whilst anecdotal evidence from colleagues at the Universities of Leicester and Southampton acknowledged that although a popular tool it was onerous, others found the ROBINS-I tool suitable for the inexperienced reviewer through well-defined domains and thorough guidance.[26] It also offered the ability to differentiate between poor quality studies in Gardener's analysis,[26] another frequent criticism of some of the shorter tools.

Whilst being acknowledged for its ease of use and a convenient scoring system,[25,27] the more established NOS has been criticized on concerns of lacking guidance,[28] inter-rater reliability,[29] and in terms of the validity of being a scoring system, which in this case rates each awarded "star" equally.[30]

In fact, several of the tools considered here have scoring systems (eg, NOS). Scoring systems are by their nature attractive for conducting subgroup analysis or meta-regression. However, Cochrane recommends against the use of scales that provide scores, as these give equal weight to each criterion being assessed without regard to their relative importance,[7] a common criticism of the Jadad score in RCTs. Berger[31] also highlights the limitations of additive scoring systems, whereby one unacceptable domain score does not relegate a study to being poor quality. This is in contrast to the ROBINS-I tool that adopts more of a multiplicative approach; a study is rated at an overall ROB equal to its

worst domain. Hence, a study is rated at serious ROB if at least one of its domains is rated as serious. The onus falls on HTA bodies to provide a clearer indication of how they wish critical appraisals to be conducted.

There is little consensus regarding which is the most appropriate critical appraisal tool to use for NRS. The different methodological quality and variability in reporting of NRS make consistent assessment of ROB difficult.[32]

Whilst HTA bodies do recommend the use of a validated critical appraisal tool, this process is undefined, and we noted considerable variability in the domains assessed in each appraisal tool. Previous research has indicated that the tools are often based on the developer's concept of research quality and that more stringent validation techniques are needed.[33]

We recommend that, to obtain the most meaningful indication of bias, reviewers should select the most appropriate NRS-specific appraisal tool for the studies identified by their systematic reviews. To comment on the appropriateness of any one NRS tool is beyond the scope of this review; however, several research groups have investigated this question. For example, the Agency for Healthcare Research and Quality recommends that the critical appraisal instrument chosen should have proven validity and transparency in how the assessments are made, be specific to the study design, and should not implement an overall composite score.[3] The use of tools designed to assess bias in RCTs should be avoided as the biases specific to NRS will not be adequately assessed by such tools. As shown by our study, appraisal tools may be specific to a particular NRS design, and some may be applicable only to prospective designs (eg, ISPOR) or to comparative studies (eg, ROBINS-I). Indeed, because of significant differences in study design, more than one tool may be necessary for any given systematic review. The interpretation and ease of use of the tool will also come into consideration: some of the tools described require considerable knowledge of study design (JBI MAStARI), whilst others have extensive domain descriptions that may take a considerable amount of time to work through (eg, ISPOR and ROBINS-I). The Cochrane ROB tool for RCTs has a simple "traffic light" system to display an overall summary of the ROB in included studies; unfortunately, none of the tools identified for NRS offer such a straightforward but effective output. Appraisal tools that exceed the 7-item length in the Cochrane ROB tool for RCTs may be more difficult to interpret, time consuming to complete, and may be more prone to error to be used in meaningful discussions around study quality. Assessment of the quality of NRS is becoming more and more important as such evidence, together with real-world data, plays an increasingly important role in health care decision making.

5 | CONCLUSION

There is little consensus around the most appropriate critical appraisal tool to use for NRS. The authors believe that whilst no single tool exists that can be used to assess ROB across all study designs, a consensus on a small suite of tools that could be used for NRS, dependent on design of studies in the review, would be preferable to no guidance. Researchers should select the most appropriate appraisal tool for the review based upon study design, focusing on those tools specific to NRS, and practical considerations such as the size of their review to provide the most meaningful indication of bias associated with included studies, whilst keeping abreast of ongoing research in this area.

ACKNOWLEDGEMENT

We thank Moira Hudson, Senior Medical Editor at ICON Health Economics, for editorial support.

CONFLICT OF INTEREST

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

All authors contributed to the study design. J.M.Q., J.C.T., and N.H. wrote the first draft, and D.A.S. provided critical review and input. All authors contributed to revisions of the manuscript and take public responsibility for its content. All authors approved the final manuscript.

AUTHORS' INFORMATION

All authors are employees of ICON Health Economics, 100 Park Drive, Milton Park, Abingdon, OX14 4RY, UK.

ORCID

Joan M. Quigley http://orcid.org/0000-0001-8658-7364

REFERENCES

1. Garrison LP Jr, Neumann PJ, Erickson P, Marshall D, Mullins CD. Using real-world data for coverage and payment decisions: the ISPOR Real-World Data Task Force report. *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research*. 2007;10(5):326-335.
2. ABPI. The vision for real world data—harnessing the opportunities in the UK. *White Paper http://wwwabpiorguk/our-work/library/industry/documents/vision-for-real-world-datapdf*. 2011.
3. AHRQ. Role of single group studies in agency for healthcare research and quality comparative effectiveness reviews. *Research White Paper http://wwweffectivehealthcareahrqgov/ehc/products/501/1389/White-Paper-Role-of-single-group-studies-1-23-13pdf*. 2013.
4. Hartwell D, Cooper K, Frampton GK, Baxter L, Loveman E. The clinical effectiveness and cost-effectiveness of peginterferon alfa and ribavirin for the treatment of chronic hepatitis C in children and young people: a systematic review and economic evaluation. *Health Technol Assess (Winchester, England*. 2014;18(65):1-202.
5. EMA. Product information 01/12/2015 Scenesse -EMEA/H/C/002548 -IB/0006, 2015.
6. Sterne JA, Hernan MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ*. 2016;355:i4919.
7. Higgins JP. *Cochrane Handbook for Systematic Reviews of Interventions*. v.5.1.0 ed2011; 2011.
8. Centre for reviews and dissemination. *Systematic Reviews: CRD's Guidance for Undertaking Reviews in Healthcare*. 3rd ed. UK: University of York; 2009.
9. SIGN. http://www.sign.ac.uk/. Accessed December 28, 2017.
10. Jorgensen L, Paludan-Muller AS, Laursen DR, et al. Evaluation of the Cochrane tool for assessing risk of bias in randomized clinical trials: overview of published comments and analysis of user practice in Cochrane and non-Cochrane reviews. *Systematic reviews*. 2016;5(1):80.

11. Jordan VM, Lensen SF, Farquhar CM. There were large discrepancies in risk of bias tool judgments when a randomized controlled trial appeared in more than one systematic review. *J Clin Epidemiol*. 2017;81:72-76.
12. Scottish Intercollegiate Guidelines Network. Methodology checklist 2: randomised controlled trials. Available from: http://www.sign.ac.uk/checklists-and-notes.html. 2015.
13. Makady A, Ham RT, de Boer A, Hillege H, Klungel O, Policies GW. For use of real-world data in health technology assessment (HTA): a comparative study of six HTA agencies. *Value in health: the journal of the International Society for Pharmacoeconomics and Outcomes Research*. 2017;20(4):520-532.
14. CADATH. Canadian Agency for Drugs and Technologies in Health 2017; https://www.cadth.ca/.
15. NICE. Guide to the methods of technology appraisal 2013. 2016.
16. Pharmaceutical Benefits Advisory Committee. Guidelines for preparing submissions to the Pharmaceutical Benefits Advisory Committee. https://pbac.pbs.gov.au/. Accessed December 28, 2017.
17. Waser NK, E, Wang S, Tao R, Goring DS. *The role of real world data in single technology appraisal submissions in the United Kingdom*. Dublin: ICPE; 2016.
18. CASP. http://www.casp-uk.net/casp-tools-checklists. Accessed December 28, 2017.
19. 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 Community Health*. 1998;52(6):377-384.
20. Wells GA, Shea B, O'Connell D, Peterson J, Welch V, Losos M, Tugwell P. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp. Accessed 28/12/2017.
21. Reeves BC, Higgins JP, Ramsay C, Shea B, Tugwell P, Wells GA. An introduction to methodological issues when including non-randomised studies in systematic reviews on the effects of interventions. *Res Synth Methods*. 2013;4(1):1-11.
22. Sterne J, Higgins JPT, Elbers RG, Reeves BC, and the development group for ROBINS-I. Risk Of Bias In Non-randomized Studies of Interventions (ROBINS-I): detailed guidance. Updated 12 October 2016; http://www.riskofbias.info. Accessed 28/12/2017.
23. Moola S, Munn Z, Tufanaru C, Aromataris E, Sears K, Sfetcu R, Currie M, Qureshi R, Mattis P, Lisy K, Mu P-F. *Chapter 7: Systematic Reviews of Etiology and Risk Australia: The Joanna Briggs Institute*; 2017.
24. Slim K, Nini E, Forestier D, Kwiatkowski F, Panis Y, Chipponi J. Methodological index for non-randomized studies (MINORS): development and validation of a new instrument. *ANZ J Surg*. 2003;73(9):712-716.
25. Deeks JJ, Dinnes J, D'Amico R, et al. Evaluating non-randomised intervention studies. *Health Technol. Assess (Winchester, England*. 2003;7(27):iii-x):1-173.
26. Gardner H. ROBINS-I: my thoughts and experience. 2017; https://heidirgardner.wordpress.com/2017/04/23/robins-i-my-thoughts-and-experience/ Accessed 28/12/2017.
27. Margulis AV, Pladevall M, Riera-Guardia N, et al. Quality assessment of observational studies in a drug-safety systematic review, comparison of two tools: the Newcastle-Ottawa Scale and the RTI item bank. *Clinical epidemiology*. 2014;6:359-368.
28. Hartling L, Milne A, Hamm MP, et al. Testing the Newcastle Ottawa Scale showed low reliability between individual reviewers. *J Clin Epidemiol*. 2013;66(9):982-993.
29. Luchini C, Stubbs B, Solmi M, Veronese N. Assessing the quality of studies in meta-analyses: advantages and limitations of the Newcastle Ottawa Scale. *WORLD JOURNAL OF META-ANALYSIS*. 2017;5(4):80-84.
30. Stang A. Critical evaluation of the Newcastle-Ottawa scale for the assessment of the quality of nonrandomized studies in meta-analyses. *Eur J Epidemiol*. 2010;25(9):603-605.
31. Berger VW, Alperson SY. A general framework for the evaluation of clinical trial quality. *Rev Recent Clin Trials*. 2009;4(2):79-88.
32. Kwan J, Sandercock P. In-hospital care pathways for stroke. *Cochrane Database Syst Rev*. 2004;4:CD002924.
33. Crowe M, Sheppard L. A review of critical appraisal tools show they lack rigor: alternative tool structure is proposed. *J Clin Epidemiol*. 2011;64(1):79-89.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Quigley JM, Thompson JC, Halfpenny NJ, Scott DA. Critical appraisal of nonrandomized studies—A review of recommended and commonly used tools. *J Eval Clin Pract*. 2018;1–9. https://doi.org/10.1111/jep.12889

The Newcastle-Ottawa Scale (NOS) for Assessing the Quality of Nonrandomized Studies in Meta-Analysis

"G. Wells, B. Shea, D. O'Connell, J. Robertson, J. Peterson, V. Welch, M. Losos, P. Tugwell"

Dr Daniel Pope – Lecturer in Epidemiology
Dr Nigel Bruce – Reader in Public Health
University of Liverpool, UK

Background

- Observational studies – aetiological hypotheses (small RR but large PAF)
- Systematic review methodology (inc. meta-analysis) attractive – precise estimate when magnitude of risk is small
- Caution required (susceptibility to bias)!

Bias and Confounding

- "...thorough consideration of sources of heterogeneity between observational study results..." Egger et al, 2003

Newcastle-Ottowa Scale

"Easy, convenient tool for quality assessment of non-randomised studies"

Newcastle-Ottowa Scale

Case-Control Studies and Cohort Studies

Star system based on three domains:

1)Selection of Study Groups

2)Comparability of Groups

3)Ascertainment of exposure/ outcome

Development: Grouping Items

- Cohort studies
  - Selection of cohorts
  - Comparability of cohorts
  - Assessment of outcome
- Case-Control studies
  - Selection of case and controls
  - Comparability of cases and controls
  - Ascertainment of exposure

Development: Identifying Items

- Identify ‘high’ quality choices with a ‘star’
- A maximum of one ‘star’ for each item within the ‘Selection’ and ‘Exposure/Outcome’ categories; maximum of two ‘stars’ for ‘Comparability’

Current Development: Validity

- Face/content validity
- Criterion validity
- Construct validity
- Inter and Intra-rater Reliability

Future Development: Scoring

- Identify threshold score distinguishing between ‘good’ and ‘poor’ quality 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: Case-Control Studies

- Selection (4)
- Comparability (1)
- Exposure (3)
  - 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 eg 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 (eg 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

Newcastle-Ottawa Quality Assessment Scale: Cohort Studies

- Selection (4)
- Comparability (1)
- Outcome (3)
  - 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 eg. 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 (eg .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 of those lost) ♦
   c) follow up rate < ___% (select an adequate %) and no description of those lost
   d) no statement

Risk of Low Birth Weight and Stillbirth Associated With Indoor Air Pollution From Solid Fuel Use in Developing Countries

Pope D, Epidemiologic Reviews, 2010

Steps of a Cochrane Systematic Review

- Clearly formulated question
- Comprehensive data search
- Unbiased selection and abstraction process
- Critical appraisal of data
- Synthesis of data
- Perform sensitivity and subgroup analyses if appropriate and possible
- Prepare a structured report

Objective

- Quantify the association between exposure to indoor air pollution and low birth weight

Inclusion Criteria

- Types of studies
  - All study designs (intervention; observational)
- Population
  - Live singleton births
- Exposure
  - Any reporting of exposure to IAP (including solid fuel use etc)
- Outcomes
  - Studies reporting actual birth weight or LBW (<2500g)

Steps of a Cochrane Systematic Review

- Clearly formulated question
- Comprehensive data search
- Unbiased selection and abstraction process
- Critical appraisal of data
- Synthesis of data
- Perform sensitivity and subgroup analyses if appropriate and possible
- Prepare a structured report

Search Strategy

- Electronic Search of:
  - MEDLINE
  - EMBASE
  - Cochrane Controlled Trials Register
  - CINAHL
  - LILACS
- Other Data Sources:
  - Grey literature (PASCAL, ICP)
  - Contact with experts, review of references cited in retrieved articles

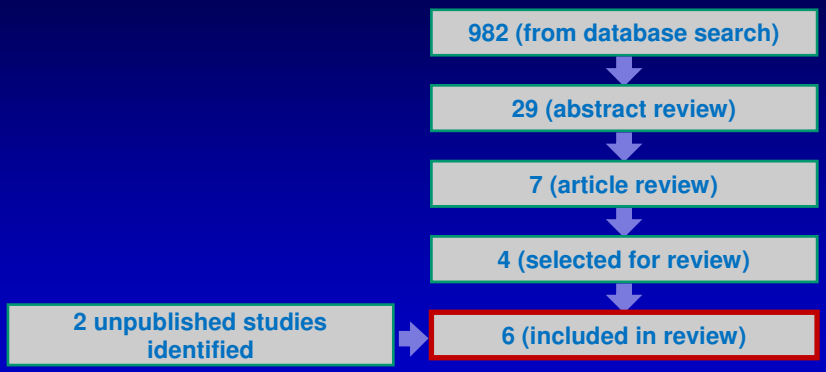
Steps of a Cochrane Systematic Review

- Clearly formulated question
- Comprehensive data search
- Unbiased selection and abstraction process
- Critical appraisal of data
- Synthesis of data
- Perform sensitivity and subgroup analyses if appropriate and possible
- Prepare a structured report

Data Extraction

- 2 independent reviewers selected studies
- 2 independent reviewers extracted data using pre-determined forms
  - study design
  - population characteristics
  - Exposure (IAP)
  - Outcomes (LBW)
  - results
- differences resolved by consensus

Results



Steps of a Cochrane Systematic Review

- Clearly formulated question
- Comprehensive data search
- Unbiased selection and abstraction process
- Critical appraisal of data
- Synthesis of data
- Perform sensitivity and subgroup analyses if appropriate and possible
- Prepare a structured report

Studies included:

- 6 studies for data extraction (from 982)
- 2 cohort
  2 cross-sectional
  1 case-control
  1 intervention study

Quality assessment:

Selection – 4 stars:

(representativeness; exposure assessment – cohort/ cross-sectional; control selection – case-control)

Comparability – 2 stars:

(adjustment for main/ additional confounders eg. active/ passive maternal smoking, gestational age, nutrition etc)

Outcome/ Exposure – 3 stars:

(adequacy of outcome (measured LBW) and exposure (indoor air pollution – measured vs self-report)

Quality assessment:

	Selection	Comparability	Outcome/ Exposure
Boy, 2002 (CS)	★★★★	★	★★★
Mavalankar, 1992 (CC)	★★★		★★★
Mishra, 2004 (CS)	★★	★	★★
Siddiqui, 2008 (C)	★★	★★	★★★
Tielsch, 2009 (C)	★★★★	★★	★★★
Thompson, 2005 (RCT)	★★★	★★	★★

Quality assessment:

	Selection	Comparability	Outcome/ Exposure
Boy, 2002 (CS)	★★★★	★	★★★
Mavalankar, 1992 (CC)	★★★		★★★
Mishra, 2004 (CS)	★★	★	★★
Siddiqui, 2008 (C)	★★	★★	★★★
Tielsch, 2009 (C)	★★★★	★★	★★★
Thompson, 2005 (RCT)	★★★	★★	★★

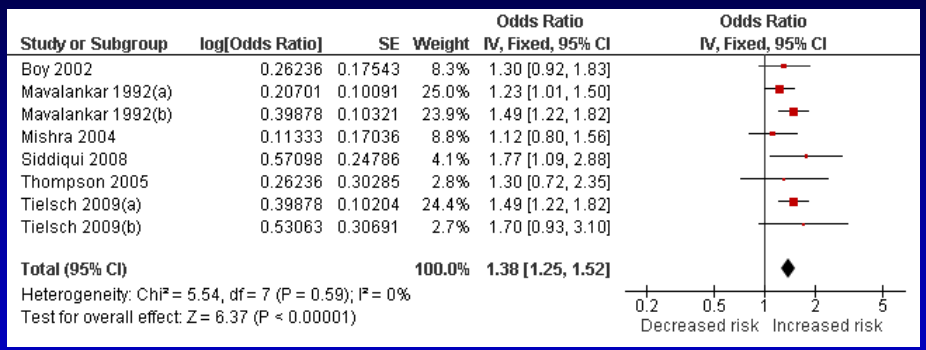
Steps of a Cochrane Systematic Review

- Clearly formulated question
- Comprehensive data search
- Unbiased selection and abstraction process
- Critical appraisal of data
- Synthesis of data
- Perform sensitivity and subgroup analyses if appropriate and possible
- Prepare a structured report

Quantification of Effects

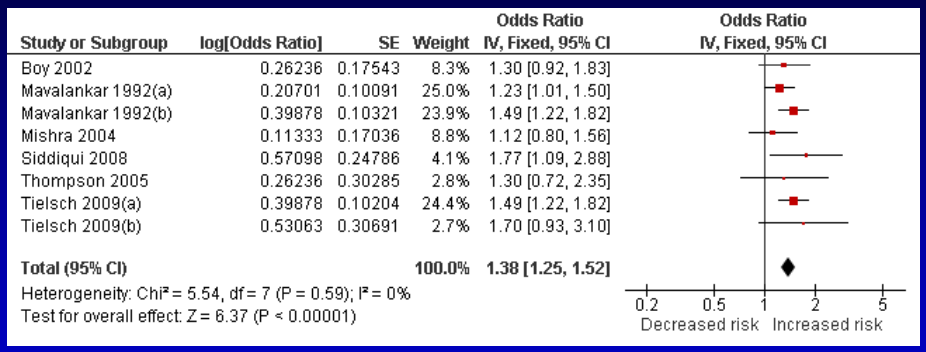
- Exposure (e.g. solid fuel vs clean fuel)
- Outcome (%LBW)
- Effect estimates (EE)
  - Relative Risk (RR)
  - Odds Ratio (OR)
- Fixed-effect meta-analysis in the absence of statistical heterogeneity

% Low Birth Weight (<2500g): 6 studies, 8 estimates



OR = 1.38 (1.25, 1.52), p<0.0001

% Low Birth Weight (<2500g): 6 studies, 8 estimates



Study or Subgroup	log[Odds Ratio]	SE	Weight	Odds Ratio IV, Fixed, 95% CI
Boy 2002	0.26236	0.17543	8.3%	1.30 [0.92, 1.83]
Mavalankar 1992(a)	0.20701	0.10091	25.0%	1.23 [1.01, 1.50]
Mavalankar 1992(b)	0.39878	0.10321	23.9%	1.49 [1.22, 1.82]
Mishra 2004	0.11333	0.17036	8.8%	1.12 [0.80, 1.56]
Siddiqui 2008	0.57098	0.24786	4.1%	1.77 [1.09, 2.88]
Thompson 2005	0.26236	0.30285	2.8%	1.30 [0.72, 2.35]
Tielsch 2009(a)	0.39878	0.10204	24.4%	1.49 [1.22, 1.82]
Tielsch 2009(b)	0.53063	0.30691	2.7%	1.70 [0.93, 3.10]
Total (95% CI)			**100.0%**	**1.38 [1.25, 1.52]**

Heterogeneity: Chi² = 5.54, df = 7 (P = 0.59); I² = 0%
Test for overall effect: Z = 6.37 (P < 0.00001)

OR = 1.38 (1.25, 1.52), p<0.0001

OR = 1.41 (1.27, 1.56) (exclude poor quality)

Interpretation Crucial:

- Exclusion from sensitivity analysis based on (i) birth weight based on self-reports (50%), (ii) no information on gestational age and (iii) unadjusted analysis

Applications:

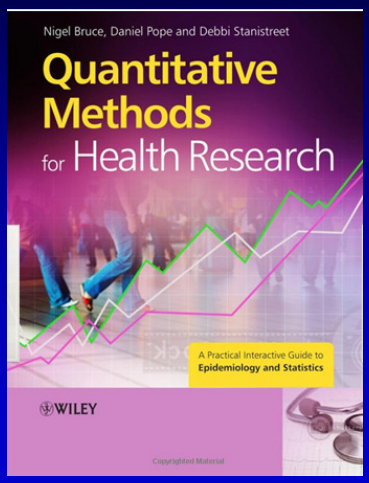
- Assess quality of nonrandomized studies
- Incorporate assessments in interpretation of meta-analytic results
- Valid, repeatable and simple
- Limitations:
  → Study Designs → Too Simplistic

The Newcastle-Ottawa Scale (NOS) for Assessing the Quality of Nonrandomized Studies in Meta-Analysis

www.lri.ca

NOS Quality Assessment Scales:
Case-control studies
Cohort studies
Manual for NOS Scales

Recommended Reading....



The Newcastle-Ottawa Scale (NOS) for assessing the quality of case-control and cohort studies

شایان مصطفایی
استادیار آمار زیستی
Shayan.mostafaei@kums.ac.ir
آبان ماه ۹۹

مقدمه

- نیوکاسل-اتاوا یک چک لیست در مطالعات مرور نظامند است که برای ارزیابی کیفیت متدولوژی(risk of bias) مطالعات مشاهدهای غیر تصادفی سازی شده اعم از مطالعات مورد-شاهدی و همگروهی/کهورت استفاده می‌شود. البته نسخه سازگار شده آن برای مطالعات مقطعی شیوع نیز موجود است.
- ابزار نیوکاسل-اتاوا حاصل همکاری بین دانشگاههای نیوکاسل در استرالیا و اتاوا در کانادا است که در سال ۲۰۰۰ ارایه شد و تاکنون حدود ۱۲ هزار ارجاع به این مقیاس داده شده است.
- این ابزار به روش دلفی و با تعامل و تکیه بر هم‌اندیشی خبرگان و بر اساس نظرات دریافتی و بازخورد کنترل شده پاسخ‌ها تهیه و اصلاح شده است.
- ارزیابی مداوم این مقیاس هنوز هم درحال انجام است.

ابزارهای دیگر ارزیابی risk of bias مطالعات همگروهی و مورد–شاهدی

- CASP
- SIGN
- NIH
- JBI

✓ Among all above mentioned tools, the NOS is the most commonly used tool nowadays

اهمیت و پیش نیازهای ارزیابی کیفیت مقالات

- ارزیابی کیفیت متدولوژی (**risk of bias**) مقالات اصیل در مطالعات مرور نظامند توسط یک مقیاس یا ابزار مناسب بسیار مورد اهمیت است.
- انتخاب یک مقیاس مناسب برای ارزیابی کیفیت مطالعات تخت تاثیر عوامل متعددی نظیر ساختار متدولوژیک مطالعات، دانش اپیدمیولوژیکال کاربر و سهولت و تعداد آیتم‌های هر مقیاس یا ابزار در این حوزه دارد.
- قابل توجه است پس از انتخاب مقیاس مناسب و آموزش افراد ارزیاب، حداقل دو ارزیاب مستقلا باید ارزیابی و بررسی متقابل برای هر مطالعه انجام دهند .معمولا شاخص توافق **Kappa** بین ارزیابان گزارش می‌شود.

دامنه آیتم‌های مقیاس NOS برای مطالعات همگروهی

- **Selection of cohorts (# items=4)**
- **Comparability of cohorts (# items=1)**
- **Assessment of outcome (# items=3)**

دامنه آیتم‌های مقیاس NOS برای مطالعات مورد-شاهدی

- **Selection of case and controls (# items=4)**
- **Comparability of cases and controls (# items=1)**
- **Ascertainment of exposure (# items=3)**

دامنه آیتم‌های مقیاس NOS برای مطالعات مورد- شاهدی و همگروهی

- **A study can be awarded a maximum of one star for each numbered item within the selection part.**
- **A study can be awarded a maximum of one star for each numbered item within the outcome/or exposure part.**
- **A maximum of two stars can be given in comparability part for each studies.**

آیتم‌های بخش انتخاب برای مطالعات مورد-شاهدی

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

آیتم‌های بخش مقایسه‌پذیری برای مطالعات مورد-شاهدی

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.)

آیتم‌های بخش مواجهه برای مطالعات مورد–شاهدی

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

آیتم‌های بخش انتخاب برای مطالعات همگروهی

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 eg 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 (eg 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

آیتم‌های بخش مقایسه‌پذیری برای مطالعات همگروهی

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.)

آیتم‌های بخش پیامد برای مطالعات همگروهی

1) Assessment of outcome
a) independent or 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

مثال موردی

Association between CD247 gene rs2056626 polymorphism and the risk of systemic sclerosis: Evidence from a systematic review and Bayesian hierarchical meta-analysis Vanaki N and et al (2019)

First Author	Year	Country	Ethnicity	NOS score	Genotyping method	Sample size		Cases			Controls		
						Case	Control	GG	GT	TT	GG	GT	TT
Dieudé	2011	French	European	8	PCR	1010	990	126	453	431	155	498	297
Wang	2014	China	Asian	7	PCR	387	523	10	82	273	7	119	397
Carmona	2016	Turkey	European	7	PCR	353	718	48	159	146	144	330	244
Abbasi	2017	Iran	Asian	7	PCR	455	455	73	208	174	65	205	185

مثال موردی

Association Study of CD226 and CD247 Genes Single Nucleotide Polymorphisms in Iranian Patients with Systemic Sclerosis

Abbasi and et.al

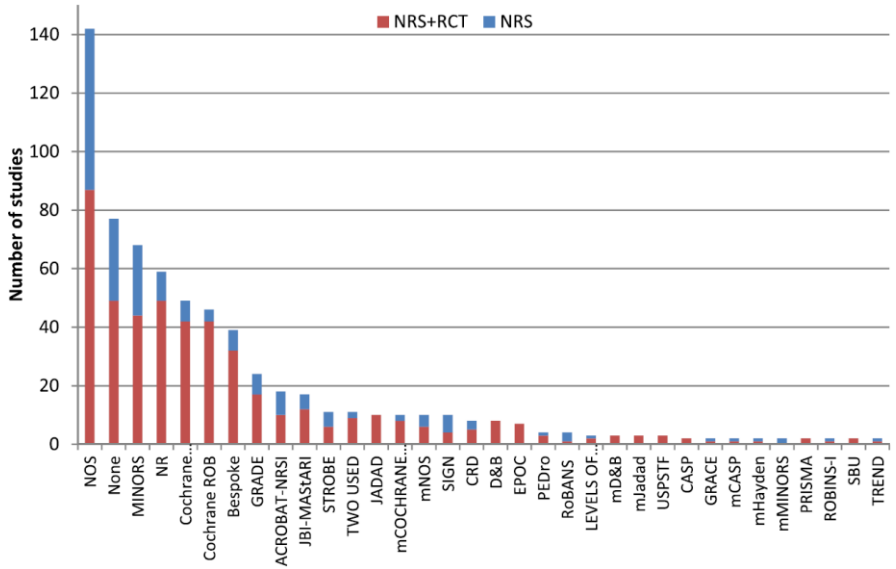
نمره یا تعداد ستاره‌های بخش انتخاب:3
نمره یا تعداد ستاره‌های بخش مقایسه پذیری:1
نمره یا تعداد ستاره‌های بخش مواجهه:3

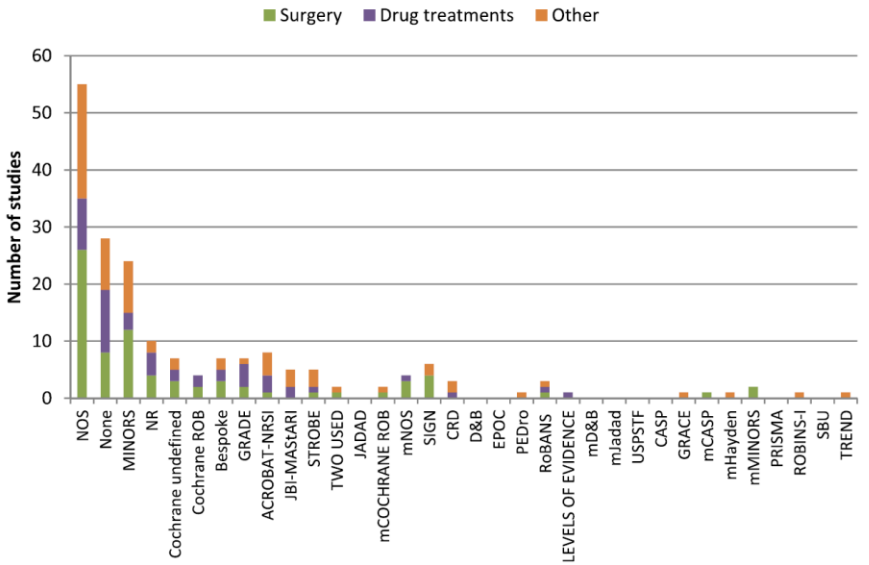
NOS vs. other tools

Critical appraisal of nonrandomized studies—A review of recommended and commonly used tools
Joan M. Quigley and et al (2018)

- Our search identified 2498 references for screening. Following fullpaper review, 686 systematic reviews that included NRS were identified and a total of 48 different critical appraisal tools were identified.
- The most commonly used critical appraisal tool for NRS was the Newcastle- Ottawa Scale for observational studies, reported in 142 of 686 (21%) of the systematic reviews. This scale allows to be modified based on a special subject.
- Whilst NOS being acknowledged for its ease of use and a convenient scoring system. The more established NOS has been criticized on concerns of lacking **standard guidance, inter-rater reliability.**

NOS vs. other tools





Reliability of NOS

Testing the Newcastle Ottawa Scale showed low reliability between individual reviewers Lisa Hartling and et al (2013)

- Objective: To assess inter-rater reliability and validity of the Newcastle Ottawa Scale (NOS) used for methodological quality assessment of cohort studies included in systematic reviews. A number of reviewers with different levels of training, type of training, and extent of experience in quality assessment and systematic reviews were included.

- Method: Two reviewers independently applied the NOS to 131 cohort studies included in eight meta-analyses
- Main result: Based on the total score, (kappa: 0.29 (0.10, 0.47)). The moderate agreement was observed in the exposure part (kappa: 0.43 (0.25, 0.61)). For all of parts of NOS, agreement was poor.
- Test–retest reliability for total score (ICC: 0.55 (95% CI 0.41 to 0.67)).

منابع

- Wells GA, Shea B, O'Connell D, Peterson J, Welch V, Losos M, Tugwell P. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses.
- Hartling L, Milne A, Hamm MP, Vandermeer B, Ansari M, Tsertsvadze A, Dryden DM. Testing the Newcastle Ottawa Scale showed low reliability between individual reviewers. Journal of clinical epidemiology. 2013 Sep 1;66(9):982-93.
- Quigley JM, Thompson JC, Halfpenny NJ, Scott DA. Critical appraisal of nonrandomized studies—A review of recommended and commonly used tools. Journal of evaluation in clinical practice. 2019 Feb;25(1):44-52.
- Ma LL, Wang YY, Yang ZH, Huang D, Weng H, Zeng XT. Methodological quality (risk of bias) assessment tools for primary and secondary medical studies: what are they and which is better?. Military Medical Research. 2020 Dec;7(1):1-1.
- Oremus M, Oremus C, Hall GB, McKinnon MC, ECT & Cognition Systematic Review Team. Inter-rater and test–retest reliability of quality assessments by novice student raters using the Jadad and Newcastle–Ottawa Scales. BMJ open. 2012 Jan 1;2(4).


Thank You!

ELSEVIER

Journal of Clinical Epidemiology 66 (2013) 982−993

Journal of Clinical Epidemiology

Testing the Newcastle Ottawa Scale showed low reliability between individual reviewers

Lisa Hartling[a,*], Andrea Milne[a], Michele P. Hamm[a], Ben Vandermeer[a], Mohammed Ansari[b], Alexander Tsertsvadze[c], Donna M. Dryden[a]

[a]*Department of Pediatrics, Alberta Research Centre for Health Evidence and the University of Alberta Evidence-based Practice Center, University of Alberta, 4-472 Edmonton Clinic Health Academy, 11405-87 Avenue, Edmonton, Alberta, Canada T5G 1C9*

[b]*Clinical Epidemiology Program, University of Ottawa Evidence-based Practice Center, The Ottawa Methods Centre, The Ottawa Hospital Research Institute, Ottawa, Ontario, Canada*

[c]*University of Ottawa Evidence-based Practice Center and Centre for Practice-Changing Research, The Ottawa Hospital Research Institute, Ottawa, Ontario, Canada*

Accepted 20 March 2013; Published online 16 May 2013

Abstract

Objectives: To assess inter-rater reliability and validity of the Newcastle Ottawa Scale (NOS) used for methodological quality assessment of cohort studies included in systematic reviews.

Study Design and Setting: Two reviewers independently applied the NOS to 131 cohort studies included in eight meta-analyses. Inter-rater reliability was calculated using kappa (κ) statistics. To assess validity, within each meta-analysis, we generated a ratio of pooled estimates for each quality domain. Using a random-effects model, the ratios of odds ratios for each meta-analysis were combined to give an overall estimate of differences in effect estimates.

Results: Inter-rater reliability varied from substantial for *length of follow-up* (κ = 0.68, 95% confidence interval [CI] = 0.47, 0.89) to poor for *selection of the nonexposed cohort* and *demonstration that the outcome was not present at the outset of the study* (κ = −0.03, 95% CI = −0.06, 0.00; κ = −0.06, 95% CI = −0.20, 0.07). Reliability for overall score was fair (κ = 0.29, 95% CI = 0.10, 0.47). In general, reviewers found the tool difficult to use and the decision rules vague even with additional information provided as part of this study. We found no association between individual items or overall score and effect estimates.

Conclusion: Variable agreement and lack of evidence that the NOS can identify studies with biased results underscore the need for revisions and more detailed guidance for systematic reviewers using the NOS.

Keywords: Methodological quality; Internal validity; Reliability; Validity; Systematic reviews; Cohort studies

1. Introduction

The internal validity of a study reflects the extent to which the design and conduct of the study have minimized the impact of bias [1]. One of the key steps in a systematic review is the assessment of internal validity (or risk of bias, RoB) of all studies included for evidence synthesis. This assessment serves to identify the strengths and limitations of the included studies; investigate and explain heterogeneity of findings across a priori defined subgroups of studies based on RoB; and grade the quality or strength of evidence for a given outcome.

With the increase in the number of published systematic reviews [2] and development of systematic review methodology over the past 15 years [1], close attention has been paid to methods of assessing internal validity of individual primary studies. Until recently, this has been referred to as ''quality assessment'' or ''assessment of methodological quality'' [1]. In this context, ''quality'' refers to ''the confidence that the trial design, conduct, and analysis has minimized biases in its treatment comparisons'' [3]. To facilitate the assessment of methodological quality, a plethora of tools has emerged [3−6]. Some of these tools are applicable to specific study designs, whereas other more generic tools may be applied to more than one design. The tools

Funding disclosure and disclaimer: This manuscript is based on a project conducted by the University of Alberta Evidence-based Practice Center under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (Contract No. 290−2007−10021). The findings and conclusions in this manuscript are those of the authors, who are responsible for its contents; the findings and conclusions do not necessarily represent the views of AHRQ. No statement in this manuscript should be construed as an official position of AHRQ or of the US Department of Health and Human Services.

* Corresponding author. Tel.: 780-492-6124; fax: 780-248-5627.
E-mail address: hartling@ualberta.ca (L. Hartling).

0895-4356/$ - see front matter
http://dx.doi.org/10.1016/j.jclinepi.2013.03.003

What is new?

- Inter-rater reliability between reviewers on the Newcastle Ottawa Scale (NOS) ranged from poor to substantial but was poor or fair for most domains.
- No association was found between individual quality domains or overall quality score and effect estimates.
- These findings underscore the need for revisions and more detailed guidance to apply the NOS in systematic reviews.

usually incorporate items associated with bias (e.g., blinding, baseline comparability of study groups) and items related to reporting (e.g., was the study population described, was a sample size calculation performed) [1].

There is a need for inter-rater reliability testing of quality assessment tools to enhance consistency in their application and interpretation across different systematic reviews. Furthermore, validity testing is essential to ensure that the tools being used can identify studies with biased results. Finally, there is a need to determine inter-rater reliability and validity to support the use of individual tools that are recommended by those developing methods for systematic reviews.

We undertook this project to assess the reliability and validity of the Newcastle Ottawa Scale (NOS). The NOS is a quality assessment tool for use on nonrandomized studies included in systematic reviews, specifically cohort and case−control studies. The tool was produced by the combined efforts of the Universities of Newcastle, Australia, and Ottawa, Canada [7], and was first reported at the Third Symposium for Systematic Reviews in Oxford, United Kingdom, in 2000 [8]. It has been endorsed for use in systematic reviews of nonrandomized studies by The Cochrane Collaboration [1].

The NOS includes separate assessment criteria for case−control and cohort studies covering the following domains: the selection of participants, comparability of study groups, and the ascertainment of exposure (for case−control studies) or outcome of interest (for cohort studies). A star rating system is used to indicate the quality of a study, with a maximum of nine stars [8]. Each criterion receives a single star if appropriate methods have been reported. The selection domain is subdivided to evaluate the selection of the exposed and nonexposed cohorts, the ascertainment of exposure, and whether the study demonstrated that the outcome of interest was not present at the start of the study. Comparability is the only category that may receive two stars: one if the most important confounders have been adjusted for in the analysis and a second star if any other adjustments were made. Outcome of interest is made up of three questions: the appropriateness of the methods used to evaluate the outcome, the length of follow-up, and the degree of the loss to follow-up [7].

The developers of the NOS have examined face and criterion validity, inter-rater reliability, and evaluator burden for the NOS. Face validity has been evaluated as strong by comparing each individual assessment item to their stem question. Criterion validity has shown a strong agreement with the Downs and Black assessment tool [9] on a series of 10 cohort studies evaluating hormone replacement therapy in breast cancer, with an intraclass correlation coefficient (ICC) of 0.88. Inter-rater reliability for the NOS on cohort studies was high with an ICC of 0.94. Evaluator burden, as assessed by the time required to complete the NOS evaluation, was shown to be significantly less than the Downs and Black tool ($P < 0.001$) [10]. The authors state that further assessment of the construct validity and the relationship between the external criterion of the NOS and its internal structures are under consideration [7]. These studies have been presented as abstracts.

The objectives of this study were to further assess the reliability of the NOS for cohort studies between individual raters and assess the validity of the NOS by examining whether effect estimates vary according to quality.

2. Methods

This article is part of a larger technical report conducted for the Agency for Healthcare Research and Quality (AHRQ). We followed a protocol that was developed a priori with input from experts in the field. Further details on methodology and results are available in the technical report (http://effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/).

2.1. Study selection

We used an iterative approach to identify a sample of cohort studies based on meta-analyses of cohort studies. Our operational definition of a cohort study was one in which individuals are grouped according to exposure status at baseline (exposed or unexposed) and are followed over time to determine if the development of the outcome of interest is different in the exposed and unexposed groups. Data may be collected prospectively or retrospectively. Initially, we searched reports completed through the Evidence-based Practice Center (EPC) Program of AHRQ to identify meta-analyses of cohort studies. We found three EPC reports [11−13] including 36 cohort studies that met the inclusion criteria. We subsequently conducted searches in MEDLINE using search terms to capture systematic reviews (meta-analys?s.mp, review.pt, and search.tw), cohort studies (exp Cohort Studies/, cohort$.tw, (observation$ adj stud$).tw) and meta-analyses (exp meta-analysis/, (analysis adj3 (group$ or pool$)).tw, (forest adj plot$).mp). Results

were limited to English language studies in humans that were published in 2000 or later. We searched by year starting with the most recent and continued until we identified a sufficient number of studies.

A meta-analysis was considered eligible for inclusion if it included estimates of at least 10 cohort studies based on a dichotomous outcome showing substantial statistical heterogeneity (i.e., $I^2 > 50\%$). Previous metaepidemiological research has used a minimum sample size per meta-analysis of 5−10 studies [14,15]. This ensures that there is a sufficient pool of studies with some degree of variability in each meta-analysis to test the hypotheses. Some degree of heterogeneity is required to test whether quality, as assessed by the NOS, can differentiate studies with different effect estimates.

Conducting sample size calculations for this type of research is challenging and cannot be done using standard approaches to sample size calculations for other research designs. The parameters required for sample size calculations for research of this nature are presently unknown. Therefore, we used a pragmatic approach to determine sample size. This was based on previous studies in this area, input from methodological experts, and the availability of resources and timelines. Thus, our target sample size was determined to consist of 125 cohort studies. Initially, 144 cohort studies from eight meta-analyses were identified; however, 13 studies were not assessed because they were later determined as having a design not relevant for this research (four randomized controlled trials [RCTs]; six case series/case−controls), or they could not be retrieved (three studies). Our final sample included 131 cohort studies; a full listing is available in the technical report (http://effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/).

2.2. Quality assessments

All studies were independently assessed by two reviewers using the NOS; the two reviewers for each study were from two different EPCs. Both centers have extensive background and experience in producing systematic reviews. For the purposes of validity assessment, discrepancies were resolved through discussion to produce consensus assessments for each study.

Reviewers pilot tested the NOS on three studies [16−18] and met by teleconference to discuss any disagreements in general interpretation of the tool. Decision rules were developed to accompany existing guidance for the NOS (Appendix A). We asked clinical experts a priori to provide the minimum length of follow-up for each review question. We identified topic-specific confounders based on the report in which the meta-analysis was included. We planned for pilot testing of an additional sample of studies if the reviewers felt there were substantial differences in interpreting and applying the tool. This was not deemed necessary after the initial pilot-testing phase.

2.3. Data extraction

The outcomes and data for effect estimates were extracted from the systematic reviews and meta-analyses and were then checked against the primary studies by a single reviewer.

2.4. Data analysis

2.4.1. Reliability of the NOS

Inter-rater agreement was calculated for each domain and for overall quality assessment using weighted [19] or unweighted Cohen kappa (κ) statistics [20], as appropriate. The former was used when domains included three or more ordinal categories, whereas the latter was used when only two categories were possible. Agreement was categorized as poor ($\kappa < 0$), slight (0−0.20), fair (0.21−0.40), moderate (0.41−0.60), substantial (0.61−0.80), or almost perfect (0.81−1.0) using accepted approaches [21]. Inter-rater agreement was based on individual reviewer assessments before discussion and consensus.

2.4.2. Validity of the RoB tool

Because there is no gold standard against which the validity of the NOS assessments can be made, we operationalized construct validity as differences in treatment effect estimates for studies that met quality criteria vs. those that did not. For example, were the effect estimates different for studies that received one star for the representativeness of the exposed cohort domain compared with studies that received no star for this domain?

For the results of the individual meta-analyses, we coded endpoints consistently so that the outcome occurrence was undesired (e.g., death as opposed to survival). Within each meta-analysis, we generated a ratio of odds ratio (i.e., odds ratios for studies with and without the domain of interest or of high/low quality as assessed by the NOS). To maintain consistency, we used odds ratios to summarize all meta-analyses, even if this was not the statistic that was used in the original meta-analysis. The ratios of odds ratios for each meta-analysis were combined to give an overall estimate of differences in effect estimates using meta-analytic techniques with inverse-variance weighting and a random-effects model [22].

2.4.3. Software

Cohen and weighted kappa statistics were obtained using StatXact (version 7.0; Cytel Inc., Cambridge, MA). Meta-analysis was done both in Stata (StataCorp, College Station, TX) and Review Manager (version 5.1.5; The Cochrane Nordic Centre, Copenhagen)

3. Results

3.1. Description of reviewers

Sixteen reviewers from the two centers assessed the studies using the NOS. Individuals had varying levels of relevant training and experience with systematic reviews

in general. The length of time they had worked with their respective EPC ranged from 4 months to 10 years. Thirteen reviewers had formal training in systematic reviews. Four reviewers had a doctoral degree; 10 reviewers had a master's degree; 1 reviewer had a medical degree and master's degree; and 1 reviewer had an undergraduate degree.

3.2. Description of sample

The cohort studies were taken from eight meta-analyses in a variety of clinical areas. These are described in Table 1 [11–13,23–27].

3.3. Inter-rater reliability

Inter-rater reliability for the 131 cohort studies is presented by domain in Table 2. The item "*was the followup long enough for the outcome to occur*" had the highest level of agreement, which was considered substantial. Reliability was moderate for both *ascertainment of exposure* and *ascertainment of outcome*. Reliability was fair for *representativeness of the cohort* and slight for *comparability of cohorts* and *adequacy of followup of cohorts*. *Selection of the nonexposed cohort* and *demonstration that the outcome was not present at the outset of the study* had poor reliability. Reliability for the overall score (total number of stars) was fair (κ = 0.29, 95% confidence interval = 0.10, 0.47).

In general, the reviewers found the tool difficult to use. They found the decision rules to be vague, even with the additional information we provided as part of this study. General points that arose were whether to assess each study based on the individual report or as it related to the systematic review question. For instance, if the systematic review question was specific to a particular population, then the study population may be representative. However, the study population may not be representative of the average population in the community (first NOS item). Similarly, reviewers wanted specific guidance on whether to base assessments on the information contained in the specific study report or whether to incorporate information from other reports of the same study. For instance, in numerous cases studies, authors would refer to another publication for details on the sample or specific methods. Studies could be unnecessarily penalized if they did not incorporate other pertinent information that was available from other reports.

Reviewers found it difficult to determine the difference between some of the response options. For example, two of the response options for item 1 regarding the exposed cohort are "truly" vs. "somewhat" representative. Some reviewers questioned whether this distinction was important, as both responses garnered a star for that item; hence, there was no difference in the final score. Also with respect to the first item, reviewers were uncertain regarding what makes a population "selected." Some interpreted this to include populations with unequal representation of a certain group (e.g., 90% males, all patients had organ transplant), whereas others relied on the methods of selection (e.g., volunteers, select group such as nurses). Likewise, reviewers questioned the difference between the "structured interview" and "written self-report" categories for ascertainment of exposure. For example, researchers may use structured validated

Table 1. Description of systematic reviews/meta-analyses of cohort studies included in sample

Topic area	Description of meta-analysis	Source	Number of studies included in our sample
Breastfeeding and asthma [11]	Association between asthma risk and breastfeeding ≥3 mo for children without family history of asthma or atopy	EPC report	10
Impaired glucose tolerance and diabetes mellitus [13]	Progression to diabetes mellitus for individuals with impaired fasting glucose or impaired glucose tolerance vs. normal glucose tolerance	EPC report	17
Cardiac resynchronization therapy and all-cause mortality [12]	All-cause mortality for individuals with implantable cardiac defibrillators compared with usual medical therapy	EPC report	11
Drug-resistant tuberculosis and positive treatment outcome [26]	Association between therapeutic approaches of extensively drug-resistant tuberculosis and favorable outcomes (i.e., cure or treatment completion)	MEDLINE	13
Statins and mortality from severe infections and sepsis [27]	Mortality from any cause in patients with sepsis and/or infection using a statin for any indication compared with placebo	MEDLINE	20
Red meat intake and prostate cancer [24]	Association between consumption (high vs. low intake) of red meat and prostate cancer	MEDLINE	15
Overweight and obesity and preterm birth before 37 wk [23]	Risk of preterm birth before 37 wk in overweight and obese women compared with women of normal weight in cohort studies	MEDLINE	38
Antenatal depression and preterm birth [25]	Association between maternal depressive symptoms during pregnancy and risk of preterm birth (<37 wk' gestation)	MEDLINE	20

Abbreviation: EPC, Evidence-based Practice Center.

Table 2. Inter-rater reliability on NOS assessments, by domain

Domain	Agreement κ[a] (95% CI)	Interpretation [21]
Representativeness of the exposed cohort	0.23 (0.09, 0.41)	Fair
Selection of the nonexposed cohort	−0.03 (−0.06, 0.00)	Poor
Ascertainment of exposure	0.43 (0.25, 0.61)	Moderate
Demonstration that the outcome was not present at outset of study	−0.06 (−0.20, 0.07)	Poor
Comparability	0.18 (−0.12, 0.47)	Slight
Assessment of outcome	0.49 (0.28, 0.70)	Moderate
Length of follow-up sufficient	0.68 (0.47, 0.89)	Substantial
Adequacy of participant follow-up	0.29 (0.12, 0.46)	Fair
Total stars	0.29[a] (0.10, 0.47)	Fair

Abbreviations: NOS, Newcastle Ottawa Scale; 95% CI, 95% confidence interval.

[a] We used a weighted kappa for the total score as it assumes some ordinality in the assessment; other kappas are not weighted, that is, Cohen kappa.

surveys or questionnaires (e.g., 36-Item Short Form Health Survey), but these are completed independently by the study participant.

Reviewers were uncertain on how to assess the item on comparability. Some studies discussed testing different confounders in their models but only included the confounders that showed a significant difference in the final model. Reviewers were unsure whether to indicate that the study adequately controlled for confounding.

Reviewers questioned what some domains actually measured. For instance, whether the selection domain assesses bias in how the participants were selected, or whether it is intended to assess the applicability of the study population to the population in general.

Reviewers would have liked ''unclear'' or ''no description'' options for some items, in particular for the last item on ''adequacy of follow-up of cohorts.'' They identified an additional problem with the response categories for this item. The second response option is either a small number lost or description provided of those lost. The third option is a larger number lost and no description of those lost. However, there is no response option that includes a larger number lost *and* a description is provided (e.g., that indicates there was no imbalance between groups).

3.4. Validity

We found no association between individual NOS items or overall NOS score and effect estimates (Table 3; Appendix B). The pooled ratios of odds ratio estimates were not statistically significant for any of the NOS items.

4. Discussion

This is the first study to our knowledge that has examined inter-rater reliability and construct validity of the NOS by researchers who were not involved in the development of the tool. We found wide variation in the degree of inter-rater agreement across the domains of the NOS, ranging from poor to substantial. The domain about the length of follow-up had substantial agreement; this finding was not surprising. The domain asked ''Was the follow-up long enough for the outcome to occur?'' Given the feedback obtained a priori from clinical experts, the assessors had very specific guidance for this item. The agreement for ascertainment of exposure and assessment of outcome was moderate, suggesting that the wording and response options are reasonable. The remaining items had poor, slight, or fair agreement that may be because of the wording of the questions and available response options. Some of the disagreement is likely attributable to inadequate reporting at the study level [1]; for example, if reporting was unclear, reviewers may have made different assumptions or interpretations of the methods used. Our findings showed less inter-rater agreement compared with an assessment completed by the developers of the tool that found a high ICC (0.94) based on assessments of 10 cohort studies by two raters [9]; however, different metrics were

Table 3. Results of meta-analysis of quality items and measures of association

Domain	ROR	95% CI
Representativeness of the exposed cohort	1.01	0.85, 1.20
Selection of the nonexposed cohort	1.83	0.92, 3.64
Ascertainment of exposure	1.13	0.93, 1.37
Demonstration that the outcome was not present at the outset of study	0.72	0.49, 1.07
Comparability	0.86	0.56, 1.31
Assessment of outcome	1.04	0.79, 1.38
Length of follow-up adequate	0.84	0.55, 1.27
Adequacy of participant follow-up	0.99	0.91, 1.08

Abbreviations: ROR, ratio of odds ratios; 95% CI, 95% confidence interval.

RORs that are greater than 1 indicate that studies of higher quality had larger effect sizes on average than studies of lower quality. The RORs presented were pooled across all the eight meta-analyses that provided data for that quality item; if all studies in a meta-analysis were rated the same for a quality item, that meta-analysis did not contribute to that ROR.

used to assess inter-rater reliability, and these may not be directly comparable.

We found no association between NOS items and effect estimates using metaepidemiological methods that control for heterogeneity because of condition and intervention. This may have been because of inadequate power, nevertheless it is consistent with previous claims that ''the NOS includes problematic items with an uncertain validity'' [28]. Previous research by the tool developers showed criterion validity by comparing the NOS with another scale (Downs and Black; ICC = 0.88); however, the sample included only 10 studies in a single topic area [9].

Although our results are less than optimal, they are not necessarily surprising. First, there is no widely accepted tool for assessing quality of nonrandomized studies. Second, there is relatively less experience of the systematic review community with nonrandomized studies as many reviews focus on RCTs. Finally, quality assessment of nonrandomized studies is inherently more challenging and possibly more subjective than for randomized trials as it involves more detailed assessment of selection bias and careful consideration of confounding.

4.1. Implications for practice

The findings of this research have important implications for practice and the interpretation of evidence. The low level of agreement between reviewers puts into question the validity of quality assessments using the NOS within any given systematic review. Moreover, in measurement theory, reliability is a necessary condition for validity (i.e., without being reliable, a test cannot be valid). Systematic reviewers are urged to incorporate considerations of study quality into their results. Furthermore, integration of the Grading of Recommendations Assessment, Development and Evaluation (GRADE) tool into systematic reviews necessitates the consideration of quality assessments in rating the strength of evidence and ultimately recommendations for practice [29]. The results and interpretation of a systematic review will be misleading if they are based on flawed assessments of quality.

Our results underscore the need for reviewers and review teams to be aware of the limitations of existing tools to assess quality of cohort studies and to be transparent in the process of quality assessment. Detailed guidelines, decision rules, and transparency are needed so that readers and end users of systematic reviews can see how the tools were applied. Furthermore, pilot testing and development of review-specific guidelines and decision rules should be mandatory and reported in detail.

The NOS in its current form does not appear to provide reliable quality assessments and requires further development and more detailed guidance. The NOS was previously endorsed by The Cochrane Collaboration [1]; however, more recently, the Collaboration has proposed a modified RoB tool to be used for nonrandomized studies [30]. A new tool developed through the EPC program for quality assessment of nonrandomized studies offers another alternative [31]. Both of these tools require independent assessment for validity and reliability.

4.2. Future directions

There is a need for more detailed guidelines to apply the NOS and revisions to the tool to enhance clarity. Additional testing should occur after further revisions to the tool and when expanded guidelines are available. We have identified specific items for which clearer guidance is needed. A living database that collects examples of quality assessments and consensus from a group of experts would be a valuable contribution to this field. Individual review teams and research groups should be encouraged to begin identifying examples, and these could be compiled across programs (e.g., the EPC program) and entities (e.g., The Cochrane Bias Methods Group) and made widely accessible. Finally, consensus in this field is needed in terms of the threshold for inter-rater reliability of a measurement before it can be used for any purpose, even descriptive purposes (i.e., describing the RoB or quality of a set of studies).

4.3. Strengths and limitations

This is the first study to our knowledge that examined reliability and validity of the NOS by researchers who were not involved in developing the tool. The main limitation of the research is that the sample size (131 cohort studies) may not have provided sufficient power to detect statistically significant differences in effect estimates according to predefined categories of quality. We specifically selected meta-analyses with substantial heterogeneity to optimize our potential to see whether quality as assessed with the NOS might explain variations in effect estimates. The results may only apply to cohort studies with a direct comparison and dichotomous outcomes.

We involved a number of reviewers with different levels of training, type of training, and extent of experience in quality assessment and systematic reviews. Some of the variability or low agreement may be attributable to characteristics of the reviewers; however, our study was not designed to examine these potential sources of variability. Agreement may be higher among individuals with more direct experience or specific postgraduate training in research methods or epidemiology. Nevertheless, all reviewers had previous experience in systematic reviews and quality assessments and likely represent the range of individuals that would typically be involved in these activities when conducting a systematic review.

5. Conclusions

More specific guidance is needed to apply and interpret quality assessment tools. We identified specific items

within the NOS where agreement is low. This information provides direction for more detailed guidance. Low agreement between reviewers has implications for incorporation of quality assessments into results and grading the strength of evidence. The low agreement, combined with no evidence that the NOS is able to identify studies with biased results, underscores the need for revisions and more detailed guidance to apply the NOS in systematic reviews.

Acknowledgments

The authors gratefully acknowledge the following individuals from the University of Alberta (U of A) EPC and University of Ottawa (U of O) EPC for assisting with quality assessments: Susan Armijo Olivo (U of A), Christine Ha (U of A), Chantelle Garritty (U of O), Kristin Konnyu (U of O), Dunsi Oladel-Rabiu (U of A), Larissa Shamseer (U of O), Kavita Singh (U of O), Elizabeth Sumamo (U of A), Jennifer Tetzlaff (U of O), Lucy Turner (U of O), Fatemeh Yazdi (U of O). We thank Annabritt Chisholm from the U of A EPC for technical assistance in preparing the report. We thank Paul Shekelle and Lina Santaguida for methodological advice.

Appendix A

Decision rules for application of the Newcastle Ottawa Scale

The following coding instructions are taken from the Newcastle Ottawa Scale Web site, http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp. Text in italics indicates additional guidance for reviewers agreed on during the initial training teleconference.

Coding manual for cohort studies

Selection

1. Representativeness of the exposed cohort

Item is assessing the representativeness of exposed individuals in the community not the representativeness of the sample of women from some general population. For example, subjects derived from groups likely to contain middle class, better educated, health-oriented women are likely to be representative of postmenopausal estrogen users, although they are not representative of all women (e.g., members of a health maintenance organization [HMO]) will be a representative sample of estrogen users. Although the HMO may have an underrepresentation of ethnic groups, the poor, and poorly educated, these excluded groups are not the predominant users of estrogen).

a) Truly representative of the average in the community*
b) Somewhat representative of the average in the community*
c) Selected group of users, for example, nurses, volunteers
d) No description of the derivation of the cohort

2. Selection of the nonexposed 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 nonexposed cohort

3. Ascertainment of exposure

a) Secure record (e.g., surgical records, *medical records*)*
b) Structured interview*
c) Written self-report

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

In the case of mortality studies, outcome of interest is still the presence of a disease/incident, rather than death. That is to say that a statement of no history of disease or incident earns a star.

a) Yes*
b) No

Comparability

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

A maximum of two stars can be allotted in this category.

Either exposed and nonexposed individuals must be matched in the design, and/or confounders must be adjusted for in the analysis. Statements of no differences between groups or that differences were not statistically significant are not sufficient for establishing comparability. Note: If the relative risk for the exposure of interest is adjusted for the confounders listed, then the groups will be considered to be comparable on each variable used in the adjustment.

There may be multiple ratings for this item for different categories of exposure (e.g., ever vs. never, current vs. previous or never).

Please see the accompanying background sheet to determine what confounders are considered important for each review topic.

If the outcome/condition of interest is gender specific (i.e., depression in pregnancy), only evaluate "a" on whether the researchers controlled for age.

a) Study controls for *age/sex* (the most important factor)*
b) Study controls for any additional factor*

Outcome

1. Assessment of outcome

For some outcomes (e.g., fractured hip), reference to the medical record is sufficient to satisfy the requirement for confirmation of the fracture. This would not be adequate for vertebral fracture outcomes where reference to X-rays would be required.

a) Independent or blind assessment stated in the paper or confirmation of the outcome by reference to secure records (X-rays, medical records, etc.)*
b) Record linkage (e.g., identified through *International Classification of Diseases* codes on database records)*
c) Self-report (i.e., no reference to original medical records or X-rays to confirm the outcome)
d) No description.

2. Was follow-up long enough for outcomes to occur?

Please see the accompanying background sheet to determine what the minimum required follow-up period is for each review topic.

a) Yes*
b) No

If the follow-up period is reported with a mean and a range, and the mean is longer than the required minimum, rate it as "yes."

3. Adequacy of follow-up of cohorts

This item assesses the follow-up of the exposed and non-exposed cohorts to ensure that losses are not related to either the exposure or the outcome.

a) Complete follow-up, all subjects accounted for*
b) Subjects lost to follow-up are unlikely to introduce bias—small number lost (less than *20%*)
c) Follow-up rate *<80%* and no description of those lost
d) No description or *unclear*

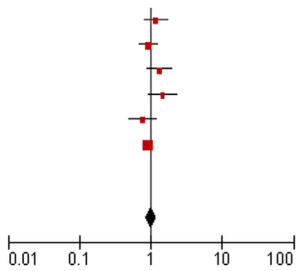
If follow-up rates vary by outcome, use the outcome included in the meta-analysis of the systematic review the article is included in.

If less than 20% of subjects were lost to follow-up, but the difference between groups is large consider downgrading to "c," especially if no reasons for difference in follow-up are provided.

Appendix B. Ratios of odds ratios for each meta-analysis and all meta-analyses combined for each quality item in the Newcastle Ottawa Scale

a) Representativeness of the exposed cohort

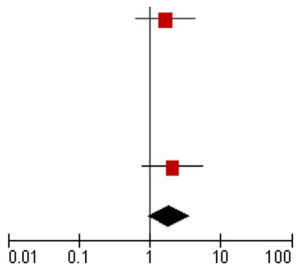
Review	Ratio of odds ratio (95% confidence interval)
1. Statins	1.16 (0.79, 1.69)
2. Prostate cancer	0.92 (0.68, 1.24)
3. Obesity	1.30 (0.87, 1.94)
4. Glucose	1.44 (0.92, 2.27)
5. Breastfeeding	0.76 (0.50, 1.15)
6. Depression	0.89 (0.76, 1.05)
7. Cardiac radiation therapy	Not estimable
Total (95% confidence interval)	1.01 (0.85, 1.20)



Heterogeneity: $Tau^2 = 0.02$; $Chi^2 = 8.38$, df = 5 ($P = 0.14$); $I^2 = 40\%$.
Test for overall effect: $Z = 0.09$ ($P = 0.93$).

b) Selection of the nonexposed cohort

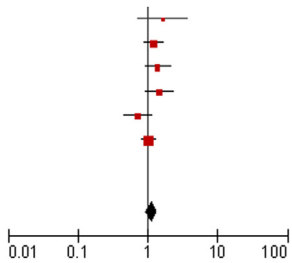
Review	Ratio of odds ratio (95% confidence interval)
1. Statins	1.64 (0.62, 4.33)
2. Prostate cancer	Not estimable
3. Obesity	Not estimable
4. Glucose	Not estimable
5. Breastfeeding	Not estimable
6. Depression	Not estimable
7. Cardiac radiation therapy	2.05 (0.77, 5.41)
Total (95% confidence interval)	1.83 (0.92, 3.64)



Heterogeneity: $Tau^2 = 0.00$; $Chi^2 = 0.10$, df = 1 ($P = 0.75$); $I^2 = 0\%$.
Test for overall effect: $Z = 1.72$ ($P = 0.08$).

c) Ascertainment of exposure

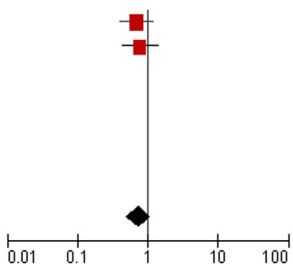
Review	Ratio of odds ratio (95% confidence interval)
1. Statins	1.62 (0.72, 3.64)
2. Prostate cancer	1.20 (0.88, 1.64)
3. Obesity	1.36 (0.89, 2.08)
4. Glucose	1.44 (0.92, 2.27)
5. Breastfeeding	0.71 (0.45, 1.13)
6. Depression	1.01 (0.81, 1.26)
7. Cardiac radiation therapy	Not estimable
Total (95% confidence interval)	1.13 (0.93, 1.37)



Heterogeneity: $Tau^2 = 0.02$; $Chi^2 = 7.57$, df = 5 ($P = 0.18$); $I^2 = 34\%$.
Test for overall effect: $Z = 1.22$ ($P = 0.22$).

d) Demonstration that the outcome was not present at the outset of study

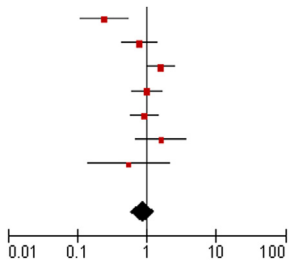
Review	Ratio of odds ratio (95% confidence interval)
1. Statins	0.69 (0.40, 1.19)
2. Prostate cancer	0.76 (0.43, 1.36)
3. Obesity	Not estimable
4. Glucose	Not estimable
5. Breastfeeding	Not estimable
6. Depression	Not estimable
7. Cardiac radiation therapy	Not estimable
Total (95% confidence interval)	0.72 (0.49, 1.07)



Heterogeneity: $Tau^2 = 0.00$; $Chi^2 = 0.06$, df = 1 ($P = 0.81$); $I^2 = 0\%$.
Test for overall effect: $Z = 1.61$ ($P = 0.11$).

e) Comparability

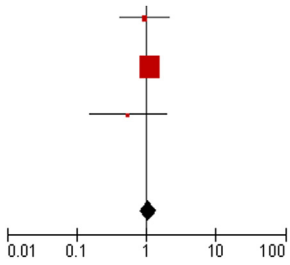
Review	Ratio of odds ratio (95% confidence interval)
1. Statins	0.24 (0.11, 0.52)
2. Prostate cancer	0.76 (0.43, 1.36)
3. Obesity	1.57 (0.97, 2.52)
4. Glucose	1.00 (0.61, 1.64)
5. Breastfeeding	0.91 (0.58, 1.43)
6. Depression	1.59 (0.68, 3.69)
7. Cardiac radiation therapy	0.54 (0.14, 2.12)
Total (95% confidence interval)	0.86 (0.56, 1.31)



Heterogeneity: $Tau^2 = 0.21$; $Chi^2 = 18.91$, df = 6 ($P = 0.004$); $I^2 = 68\%$.
Test for overall effect: $Z = 0.69$ ($P = 0.49$).

f) Assessment of outcome

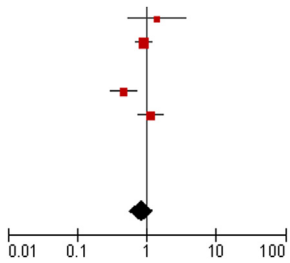
Review	Ratio of odds ratio (95% confidence interval)
1. Statins	0.93 (0.41, 2.11)
2. Prostate cancer	Not estimable
3. Obesity	1.10 (0.81, 1.50)
4. Glucose	Not estimable
5. Breastfeeding	0.53 (0.15, 1.89)
6. Depression	Not estimable
7. Cardiac radiation therapy	Not estimable
Total (95% confidence interval)	1.04 (0.79, 1.38)



Heterogeneity: $Tau^2 = 0.00$; $Chi^2 = 1.28$, df = 2 ($P = 0.53$); $I^2 = 0\%$.
Test for overall effect: $Z = 0.31$ ($P = 0.76$).

g) Length of follow-up adequate

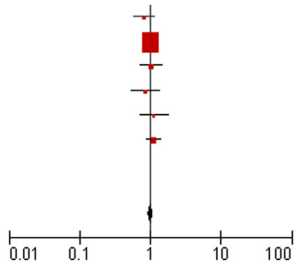
Review	Ratio of odds ratio (95% confidence interval)
1. Statins	1.38 (0.52, 3.63)
2. Prostate cancer	0.89 (0.69, 1.16)
3. Obesity	Not estimable
4. Glucose	0.45 (0.29, 0.71)
5. Breastfeeding	1.11 (0.73, 1.69)
6. Depression	Not estimable
7. Cardiac radiation therapy	Not estimable
Total (95% confidence interval)	0.84 (0.55, 1.27)



Heterogeneity: $Tau^2 = 0.12$; $Chi^2 = 10.24$, df = 3 ($P = 0.02$); $I^2 = 71\%$.
Test for overall effect: $Z = 0.85$ ($P = 0.40$).

h) Adequacy of participant follow-up

Review	Ratio of odds ratio (95% confidence interval)
1. Statins	0.79 (0.57, 1.12)
2. Prostate cancer	0.99 (0.89, 1.10)
3. Obesity	1.01 (0.71, 1.44)
4. Glucose	0.85 (0.54, 1.33)
5. Breastfeeding	1.10 (0.70, 1.75)
6. Depression	1.11 (0.89, 1.39)
7. Cardiac radiation therapy	Not estimable
Total (95% confidence interval)	0.99 (0.91, 1.08)



Heterogeneity: $Tau^2 = 0.00$; $Chi^2 = 3.28$, df = 5 ($P = 0.66$); $I^2 = 0\%$.
Test for overall effect: $Z = 0.19$ ($P = 0.85$).

References

[1] Higgins PT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions. London, UK: The Cochrane Collaboration; 2009. 5.0.2 [updated September 2009].

[2] Bastian H, Glasziou P, Chalmers I. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? PLoS Med 2010;7(9):e1000326.

[3] 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.

[4] 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.

[5] West S, King V, Carey TS, Lohr KN, McKoy N, Sutton SF, et al. Systems to rate the strength of scientific evidence. Evid Rep Technol Assess (Summ) 2002;47:1−11.

[6] Olivo SA, Macedo LG, Gadotti IC, Fuentes J, Stanton T, Magee DJ. Scales to assess the quality of randomized controlled trials: a systematic review. Phys Ther 2008;88:156−75.

[7] Wells G, Shea B, O'Connell J, Robertson J, Peterson V, Welch V, et al. The Newcastle-Ottawa scale (NOS) for assessing the quality of nonrandomised studies in meta-analysis. Available at http://www.ohri.ca/programs/clinical_epidemiology/oxfordasp. Accessed April 13, 2013.

[8] Wells G, Shea B, O'Connell J, Robertson J, Peterson V, Welch V, et al. The Newcastle-Ottawa scale (NOS) for assessing the quality of nonrandomised studies in meta-analysis [abstract]. Oxford, United Kingdom, July 3−5, 2000.

[9] 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 Community Health 1998;52:377−84.

[10] Wells G, Brodsky L, O'Connell D, Robertson J, Peterson V, Welch V, et al. Evaluation of the Newcastle-Ottawa Scale (NOS): an assessment tool for evaluating the quality of non-randomized studies [abstract]. Barcelona, Spain, October 26−31, 2003.

[11] Ip S, Chung M, Raman G, Chew P, Magula N, DeVine D, et al. Breastfeeding and maternal and infant health outcomes in developed countries. Evid Rep Technol Assess (Full Rep) 2007;153:1−186.

[12] McAlister FA, Ezekowitz J, Dryden DM, Hooton N, Vandermeer B, Friesen C, et al. Cardiac resynchronization therapy and implantable cardiac defibrillators in left ventricular systolic dysfunction. Evid Rep Technol Assess (Full Rep) 2007;152:1−199.

[13] Santaguida PL, Balion C, Hunt D, Morrison K, Gerstein H, Raina P, et al. Diagnosis, prognosis, and treatment of impaired glucose tolerance and impaired fasting glucose. Evid Rep Technol Assess 2005;128:1−11.

[14] Egger M, Juni P, Bartlett C, Holenstein F, Sterne J. How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study. Health Technol Assess 2003;7:1−76.

[15] Furukawa TA, Watanabe N, Omori IM, Montori VM, Guyatt GH. Association between unreported outcomes and effect size estimates in Cochrane meta-analyses. JAMA 2007;297:468−70.

[16] Norman R, Masters L, Milner C, Wang J, Davies M. Relative risk of conversion from normoglycaemia to impaired glucose tolerance or non-insulin dependent diabetes mellitus in polycystic ovarian syndrome. Hum Reprod 2001;16:1995−8.

[17] Wright A, Holberg C, Taussig L, Martinez F. Factors influencing the relation of infant feeding to asthma and recurrent wheeze in childhood. Thorax 2001;56(3):192−7.

[18] Arvanitakis Z, Schneider JA, Wilson RS, Bienias JL, Kelly JF, Evans DA, et al. Statins, incident Alzheimer disease, change in cognitive function, and neuropathology. Neurology 2008;70(19 Pt 2): 1795−802.

[19] Liebetrau A. Measures of association. Newbury Park, CA: Sage Publications; 1983.

[20] Agresti A. Categorical data analysis. New York, NY: John Wiley & Sons; 2002.

[21] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159−74.

[22] Sterne JA, Juni P, Schulz KF, Altman DG, Bartlett C, Egger M. Statistical methods for assessing the influence of study characteristics on treatment effects in 'meta-epidemiological' research. Stat Med 2002;21:1513−24.

[23] McDonald SD, Han Z, Mulla S, Beyene J. Overweight and obesity in mothers and risk of preterm birth and low birth weight infants: systematic review and meta-analyses. BMJ 2010;341:c3428.

[24] Alexander DD, Mink PJ, Cushing CA, Sceurman B. A review and meta-analysis of prospective studies of red and processed meat intake and prostate cancer. Nutr J 2010;9:50.

[25] Grote NK, Bridge JA, Gavin AR, Melville JL, Iyengar S, Katon WJ. A meta-analysis of depression during pregnancy and the risk of preterm birth, low birth weight, and intrauterine growth restriction. Arch Gen Psychiatry 2010;67:1012−24.

[26] Jacobson KR, Tierney DB, Jeon CY, Mitnick CD, Murray MB. Treatment outcomes among patients with extensively drug-resistant tuberculosis: systematic review and meta-analysis. Clin Infect Dis 2010;51:6−14.

[27] Janda S, Young A, Fitzgerald JM, Etminan M, Swiston J. The effect of statins on mortality from severe infections and sepsis: a systematic review and meta-analysis. J Crit Care 2010;25(4): 656−722.

[28] Stang A. Critical evaluation of the Newcastle-Ottawa scale for the assessment of the quality of nonrandomized studies in meta-analyses. Eur J Epidemiol 2010;25(9):603−5.

[29] Schunemann HJ, Brozek J, Oxman AD. GRADE handbook for grading quality of evidence and strength of recommendation. Version 3.2 [update March 2009]. The Grade Working Group. Available at http://www.gradeworkinggroup.org/index.htm; Accessed April 13, 2013.

[30] Higgins PT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions. London, UK: The Cochrane Collaboration; 2011. 5.1.0. [updated March 2011].

[31] Berkman ND, Viswanathan M. Development of a tool to evaluate the quality of non-randomized studies of interventions or exposures. Bethesda, MD, September 15, 2009. Available at http://ahrq.gov/legacy/about/annualconf09/berkman_viswanathan/berkman_viswanathan.ppt.