Abstract

Objective: It is useful for reviewers of economic evaluations to assess quality in a manner which is consistent and comprehensive. Checklists can allow this, but there are concerns about their reliability, and how they are used in practice. We aimed to describe how checklists have been used in systematic reviews of health economic evaluations.

Study Design: Meta-review with snowball sampling

Methods: We compiled a list of checklists for health economic evaluations and searched for the checklists' use in systematic reviews from 2010 – February 2018. We extracted data regarding checklists used, stated checklist function, subject area, number of reviewers and issues expressed about checklists.

Results: We found 346 systematic reviews since 2010 which used checklists to assess economic evaluations. The most common checklist in use was developed in 1996 by Drummond and Jefferson and the most common stated use of a checklist was as a quality assessment. Checklists and their use varied within subject areas. 223 reviews had more than one reviewer who used the checklist.

Conclusions: Use of checklists is inconsistent. Eighteen individual checklists have been used since 2010, many of which have been used in ways different from those originally intended, often without justification. Different systematic reviews in the same subject areas would benefit from using one checklist exclusively, using checklists as intended, and having two reviewers complete the checklist. This would increase the likelihood that results are transparent and comparable over time.

Introduction:

A checklist is commonly used in reviews to standardise the assessment (e.g. of quality or completeness) of economic evaluations.**[1]** Checklists contain a list of criteria or 'items' against which a reviewer assesses the study. Sometimes a reviewer may use a checklist to provide a score of general quality (e.g. 'poor' quality might be ascribed if only 50% of items are successfully completed).

Checklists, like the studies they assess, are of variable quality, with previous reviews raising contentions about test-retest reliability and inter-rater reliability.**[1,2]** Checklists also vary in what criteria they assess. For example, when ten common checklists were compared to a checklist developed by Evers et al.,**[3]** the proportion of similar items ranged from 32% to 95%.**[4]** There are also many checklists employed to perform the same tasks (e.g. quality assessment), with earlier checklists appearing in the early 1990s,**[5,6]** and checklists still being published recently.**[7]** In a previous review examining systematic reviews from 1990 – 2010, the authors emphasised the importance of using comparable instruments to assess quality, as few reviews used the same instrument, often finding instruments developed ad-hoc.**[8]** These issues have cast doubt on the usefulness of checklists to produce a standardised and transparent assessment of health economic evaluations. This study adds to the existing literature by answering the following questions:

1.) Which checklists are commonly in use? Is this consistent within subject areas?

2.) Why do reviewers use checklists?

3.) When using checklists, how do reviewers present their findings?

4.) When assessing quality, how many reviewers complete the assessment process?

5.) What issues do reviewers report about checklists?

Methods:

Searching for checklists

Definition and eligibility criteria

Our definition of a checklist was adopted from Walker et al. (2012),**[9]** as "any original listing of specific items that the authors recommended be addressed in the conduct or reporting of an economic evaluation." Our eligibility criteria included checklists in peer-reviewed publications between 1992 and 2015. Checklists were excluded if they were not published in English. In contrast with Walker et al.,**[9]** our study included guidelines, questionnaires or other instruments, which differ from checklists conceptually (e.g. guidelines give instruction on how to perform a study prior to publication,) if they had been used as if they were a checklist.

Search strategy and study selection

We created a database of previously published lists of checklists from two systematic reviews conducted by Walker et al. (2012)**[9]** and Moher et al. (2011).,**[4]** recommendations from the EQUATOR network,**[10]** and a table of checklists compiled by Husereau et al. (2013).**[11]** Checklists which were found during the search for reviews of economic evaluations were also added to this database.

Searching for reviews of economic evaluations

Definition and eligibility criteria

We used this database to search for articles which had cited these checklists, identifying potential reviews for analysis. Our eligibility criteria included reviews which identified themselves as systematic reviews and identified the use of at least one checklist for assessing economic evaluations. Reviews were excluded if they were published before 2010, or were published in a language other than English.

Search strategy and study selection

This review utilised a forward citation search strategy using the Web of Science via ISI database. Each checklist was searched for articles which had cited it using the Web of Science citation network function. Returned citations were then screened for eligibility based on title and abstract and catalogued using the Zotero reference manager (George Mason University, Virginia). Eligible reviews were then obtained using the online OneSearch service (The University of Western Australia, Western Australia) and screened again based on the complete text and supplementary materials.

Data extraction and synthesis

Data extraction was performed by one reviewer (RW) using Excel 2016 (Microsoft Corporation, Washington). The data extracted was as follows: checklist used, whether the checklist was modified (by modifying or adding criteria, or by changing the way it is used e.g. adding weightings for scoring purposes), the stated reason for using a checklist, whether the reviewers provided a score using the checklist, reported issues about the checklists themselves, the number of reviewers who conducted the checklist assessment, whether reviewers specified how they resolved conflicts, year of publication, journal and subject area of the review. Subject area was labelled using two methods: by Scimago Journal Category(ies) and by keywords ascribed by the reviewer (RW) by screening the review. Narrative synthesis (the attempt to organise, explore, describe and interpret study findings **[12]**) was used to identify reported issues about checklists from reviews: methods, results and discussion sections were reviewed, with issues and themes (such as discussion issues or scoring method) being listed in the author’s own words. After all issues were described, common issues and themes were synthesised.

Statistical methods

Descriptive statistical methods were used to present quantitative data, such as counts and averages. Results were calculated in Python using Jupyter notebooks, Pandas, MatPlotLib and Numpy. The Jupyter notebook showing our analysis is available in Supplement A.

Results:

Study selection

We identified 38 checklists from the period of 1992 **[5,6]** to 2015.**[7]** The reference list for these checklists can be found in Supplement B. These checklists yielded a total of 6,179 forward citations, of which 638 of the citations were found to be relevant based on title and abstract screening. After screening the full content of these 638 review papers, 346 were found to be relevant to our review. Figure 1 illustrates this process.

Description of the identified reviews

Of the 346 reviews identified, 24 reviews (7%) were from 2010, 21 reviews (6%) were from 2011, 37 reviews (11%) were from 2012, 30 reviews (9%) were from 2013, 44 reviews (13%) were from 2014, 62 reviews (18%) were from 2015, 55 reviews (16%) were from 2016, 71 reviews (20%) were from 2017 and three reviews were published in January of 2018 (<1%). The most common journals where reviews were published were Health Technology Assessment (n=67, 19%), followed by Pharmacoeconomics (n=32, 9%), and PLoS One (n=21, 6%). There were a total of 173 unique journals which published reviews using economic evaluation checklists. Table 1 describes characteristics of the identified reviews.

Which checklists are commonly in use? Was this consistent in subject areas?

In total, there were 439 uses of 18 unique checklists. The most common checklist used was the 36-item Drummond and Jefferson checklist from the British Medical Journal (hereafter referred to as the BMJ checklist)**[13]** which was used in 117 (30%) reviews in total. The second most common checklist was the CHEC-list developed by Evers et al., used in 77 (18%) reviews.**[3]**

Following these were the Philips checklist (n=59, 13%),**[14]** the Consolidated Health Economic Evaluation Reporting Standards (CHEERS) checklist (n=59, 13%),**[11]** and the Quality of Health Economic Studies (QHES) checklist (n=58, 13%).**[15]** The Drummond Ten Point checklist was also commonly cited (n=41, 9%).**[16]** The CHEERS checklist has experienced the largest increase in use since its development, as illustrated in Figure 2.

In 90 reviews (26%) there were two or more checklists used. The three most common pairings of checklists were the BMJ and Philips checklist (n=18, 20%), the CHEC and Philips checklist (n=14, 16%) and the BMJ and CHEC checklists (n=11, 12%). Furthermore, 82 reviews (24%) included checklists which had been modified from their original design. Commonly, modifications were ad-hoc (e.g. the mQHES modification by Nuckols et al.**[17, 18])**, but cited modifications e.g. the 'La Torre' modification to the BMJ checklist (which adds a weighted scoring system)**[19]** occurred as well.

Regardless of method, the checklists which were used varied substantially within subject areas. Supplement C covers these results comprehensively.

Why were checklists used?

There were four stated uses of checklists in the identified reviews: an assessment of quality, a method of data extraction, a 'risk of bias' assessment and a guide for eligibility criteria. Checklists were used for quality assessment in 94% of reviews, data extraction in 2.3%, risk of bias in 2.3% and eligibility criteria in 1.4%.

When using checklists, how do reviewers present their findings?

199 reviews (57.7%) presented the results of their quality assessment as a quantitative score (e.g. 91/100 or 85%). When the QHES checklist was removed from these (the QHES checklist intends reviewers to score studies), 141 reviews (49.1%) still provided a score. Where reviews scored economic evaluations and provided a threshold that indicated ‘high quality’, the threshold varied. The minimum reported cut-off for an evaluation to be considered high quality was 63%, and the maximum was 94%.

Quality assessments were presented in five ways: a quality rating (e.g. ‘poor’ or ‘good’), a study score (e.g. Study 1 scored 9/10), a global item score (e.g. the average score for checklist item 3 was 75%), an individual item score (e.g. item 3 for Study 1 scored was completed correctly) and a domain specific score (e.g. items regarding sensitivity analyses were completed 80% of the time). Figure 3 illustrates these presentations.

When assessing quality, how many reviewers complete the assessment process?

For reviews which assessed quality, risk of bias, or used a checklist for data extraction (n=341), the number of reviewers using the checklist varied. 62 reviews (18%) specified that one reviewer used the checklist, and 223 reviews (64.6%) specified more than one reviewer, with the maximum amount of reviewers being eight (two teams of four reviewers independently assessed each study). The amount of reviewers was not specified in 60 (17.4%) reviews. Where there was more than one reviewer, 165 (74%) reviews discussed how reviewers would resolve disagreements when using a checklist.

What issues do reviewers have when using checklists?

In the discussion section of the identified reviews, there were several common reported issues with checklists. Issues were raised 83 times, which could be grouped into 6 categories. The most common issue raised regarded the subjective nature of checklists (39.8% of issues raised). 22.9% of issues regarded the difficulty in use of checklists. 16.9% of issues highlighted that scores do not imply quality. 15.7% of issues regarded checklists not being comprehensive enough. Other issues highlighted the number of checklists to choose from, and that checklist use did not affect conclusions drawn from the review. Issues were most commonly raised in reviews using the QHES checklist, with issues being discussed in 20 publications (34.5% of QHES publications). Supplement C provides an overview of issues, as well as examples of such issues.

Discussion:

Findings and fit with general knowledge

Which checklists are commonly in use?

We identified a total of 346 reviews between 2010 and 2018 which used checklists. The most comparable study is Hutter et al.**[8]** who documented the use of checklists in reviews from 1990-2010, finding 42 reviews published between 1990-2001 (based on data provided by Jefferson et al.),**[20]** and 34 published between 2002-2010.**[8]** Unsurprisingly there has been an increase in the amount of published economic evaluations since 2010.**[21]** Our approach differed from Hutter et al.'s in that we included any systematic review which used a checklist, as we wanted to assess the use of checklists (such as in partial economic evaluations and cost-of-illness studies), rather than comparing and identifying the criteria used to assess quality in full economic evaluations.**[8]**

We found 18 unique checklists used with the most common being the BMJ checklist, the Philips checklist, the QHES checklist, the CHEC checklist and the CHEERS checklist. Hutter et al. found the majority of checklists were developed ad-hoc, with two reviews using the QHES checklist, four using the BMJ checklist, and five using a checklist by Gold et al.**[22]** There is a continuing preference amongst reviewers for the BMJ checklist and the QHES checklist, despite introduction of alternatives, although reviews using the CHEERS checklist have increased substantially since its development.**[11]** It also appears the use of ad-hoc checklists is decreasing, however our methodology looked for published checklists, and therefore ad-hoc checklists may have not been captured.

We can contrast our findings with the results from a review by Walker et al. (2012).**[9]** This review presented the number of times ten identified checklists had been cited, giving "an indication of the use and importance of each checklist".**[9]** They found that a checklist by Russell et al.**[23]** was the most frequently cited, followed by Weinstein et al.,**[24]** followed by the BMJ checklist. Our findings do not necessarily contradict this. We did not assess citation count, but instead measured checklist use explicitly. There are several issues with using citation count as a proxy for quality or importance, which have been discussed elsewhere.**[25-27]**

Why do reviewers employ checklists?

We found that the majority of identified reviews use a checklist as an 'assessment of quality' although these reviews did not commonly mention what quality (e.g. methodological quality, reporting quality) was being assessed by the checklist. This issue is less important with some checklists, such as the QHES checklist, which states that its function is to assess methodological quality, but for checklists such as the CHEERS checklist, uncertainty regarding whether it is assessing methodological quality or reporting quality (or 'completedness') has previously been acknowledged.**[28]** It is also interesting to note that some reviews have used checklists as guides for eligibility criteria. The exclusion of studies based on an arbitrary score may well be a source of bias for these reviews, especially where only one reviewer performs the assessment.

When using checklists, do reviewers score economic evaluations?

199 (57.7%) reviews presented a score for their quality assessments, and 141 (49.1%) reviews presented a score when the QHES checklist was excluded from the results. This is significant, because most checklists do not mention, or advise against providing a score to indicate quality. Scores tend to oversimplify a quality assessment e.g. it is unlikely that a study which scored 75%

for each item has the same quality as a study which omitted 25% of all items, although their score would be the same. A separate but relevant issue with scoring is the application of arbitrary cut-offs which indicate quality. Our review found cut-offs which indicated ‘high quality’ ranged from 63% to 94%. Presumably an evaluation which scores 63% is not the same as one which 94%, regardless of how quality is defined. We would advise against providing arbitrary cut-offs in the future.

We also found that there were several common methods of presenting the results of a quality assessment. Reviews tended to present quality ratings, global or individual item scores, study scores, domain specific scores, or some combination thereof. Whilst acknowledging possible limitations imposed by journals, we find the most useful presentation of quality assessments is a matrix with one axis showing each checklist criterion, and the other axis presenting each economic evaluation. This allows researchers and policy-makers the opportunity to aggregate and score (if appropriate) evaluations as they choose.

When assessing quality, how many reviewers complete the assessment process?

Gerkens et al. (2008)**[2]** found that the reviewer significantly influenced the results of the checklist score and thus, recommended that at least two assessors score each study.**[2]** We found that 122 reviews did not specify the number of reviewers, or only used one reviewer to assess quality. Of the 223 reviews which used more than one reviewer to assess checklists, 58 did not specify how disputes concerning the items were resolved. These findings suggest that many reviews using a checklist are presenting information which may have been different if another reviewer had been involved.

What issues do reviewers have with checklists?

Most reported issues in reviews pertained to the inherent subjectivity of the checklist being used, an issue which has been commented on in previous publications.**[1,2]** Other issues pertained to how items were assessed, or the lack of comprehensiveness of checklists themselves.

Issues were most commonly brought up regarding the QHES checklist, but this should not be assumed to be indicative of the QHES checklist being more problematic than other checklists. The QHES checklist has been validated and test-retest reliability have been repeatedly demonstrated.**[2,15**,**29,30]** The reason for more issues being brought up for the QHES checklist may be that the use of QHES to perform a quality assessment is well-described, which means reviewers have a clearer framework on which to base their contentions.

Limitations and assumptions

Search strategy

There were several limitations with this study which are important to note. Firstly, we may have been limited by our search strategy. Rather than adopt the search strategy of a systematic review, we chose to compile a list of checklists from previous reviews and guidelines, and checklists identified during our search process. Checklists have been identified by reviews previously,**[9,11]** or are included in lists to guide reviewers, such as the EQUATOR network reporting guidelines.**[10]** It seemed reasonable to use these as the basis of a search strategy which could be added to as these were systematic reviews themselves and were published within a few years of our study.**[4,9]** It is possible that checklists have been missed as a result of this strategy (especially checklists which were developed and used in the same publication, without reference to checklists we identified), but we do not believe this would have affected our conclusions as checklists referenced in previous reviews and guidelines yielded larger search results for reviews than checklists which were discovered during the search process itself.

Checklist definition and citation

There may have been instances where the incorrect checklist was cited. This is most relevant to the BMJ checklist and the Drummond Ten Point checklist, where one reference was cited (e.g. the BMJ checklist article) but another was used (e.g. the Drummond Ten Point checklist). We don't believe this would have affected the variability found substantially.

We were unable to find forward citations for the NICE reference case,**[31]** due to lack of a standardised citation. Despite finding uses of it as a checklist, attempts to search for other citations using Web of Science yielded no results. Furthermore, there was one instance where we could not find an appropriate publication for an identified checklist (see Supplement B), and thus were not able to identify other uses of it. We believe these may have reduced the final number of reviews we analysed.

Conclusion and recommendations

If one of the uses of a checklist is to standardise assessment, it must be assumed that checklist use is itself standard, and this is not currently the case. Our findings have demonstrated that there is variability in both the checklists used to assess economic evaluations, and the ways in which they are used. An assessment which has been produced by a checklist is not the only factor of importance for decision-makers, but as it stands the variability of checklists does limit the ability to provide transparent and consistent results.

As such, we make the following three recommendations to authors of reviews which use a checklist in relation to health economic evaluations:

1.) Within one's subject area, identify which checklists have been used previously. It would be more consistent to assess quality using the same checklist. It would also be more useful to use a checklist which has been validated, and is commonly used elsewhere. We note to those using the BMJ and Drummond checklists that in reviews where Drummond himself is an author, the CHEERS checklist is now used.**[32]**

2.) A checklist should be used by two reviewers at least, and there should be a mechanism in place to resolve disputes of reviewers prior to using the checklist in the review. This decreases the chance that conclusions drawn from results are biased.

3.) A checklist should serve its intended purpose. Any modification, including using a checklist to provide a score, should be made clear and be thoroughly justified.

References

1. Langer A. A framework for assessing Health Economic Evaluation (HEE) quality appraisal instruments. BMC Health Serv Res. 2012;12:253.

2. Gerkens S, Crott R, Cleemput I, et al. Comparison of three instruments assessing the quality of economic evaluations: a practical exercise on economic evaluations of the surgical treatment of obesity. Int J Technol Assess Health Care. 2008;24(3):318–25.

3. Evers S, Goossens M, de Vet H, van Tulder M, Ament A. Criteria list for assessment of methodological quality of economic evaluations: consensus on health economic criteria. Int J Technol Assess Health Care. 2005;21(2):240–5.

4. Moher D, Weeks L, Ocampo M, et al. Describing reporting guidelines for health research: a systematic review. J Clin Epidemiol. 2011;64(7):718–42.

5. Adams ME, McCall NT, Gray DT, Orza MJ, Chalmers TC. Economic analysis in randomized control trials. Med Care. 1992;30(3):231–43.

6. Gerard K. Cost-utility in practice: a policy maker's guide to the state of the art. Health Policy Amst Neth. 1992;21(3):249–79.

7. Ramsey SD, Willke RJ, Glick H, et al. Cost-effectiveness analysis alongside clinical trials ii-an ISPOR good research practices task force report. Value Health. 2015;18(2):161–72.

8. Hutter M-F, Rodríguez-Ibeas R, Antonanzas F. Methodological reviews of economic evaluations in health care: what do they target? Eur J Health Econ. 2014;15(8):829–40.

9. Walker DG, Wilson RF, Sharma R, Bridges J, et al. Best practices for conducting economic evaluations in health care: a systematic review of quality assessment tools. Rockville: Agency for Healthcare Research and Quality, 2012

10. EQUATOR network. Search for reporting guidelines – Economic evaluations reporting guidelines for economic evaluations. Available from: https://www.equator-network.org/reporting-guidelines/. [Accessed July 25, 2017].

11. Husereau D, Drummond M, Petrou S, et al. Consolidated Health Economic Evaluation Reporting Standards (CHEERS)--explanation and elaboration: a report of the ISPOR Health Economic Evaluation Publication Guidelines Good Reporting Practices Task Force. Value Health. 2013;16(2):231–50.

12. Pope C, Mays N, Popay J. Synthesising qualitative and quantitative health evidence: a guide to methods: a guide to methods. New York: McGraw-Hill Education, 2007.

13. Drummond MF, Jefferson TO. Guidelines for authors and peer reviewers of economic submissions to the. BMJ. 1996;313(7052):275–83.

14. Philips Z, Bojke L, Sculpher M, Claxton K, Golder S. Good practice guidelines for decision-analytic modelling in health technology assessment: a review and consolidation of quality assessment. PharmacoEconomics. 2006;24(4):355–71.

15. Chiou C-F, Hay JW, Wallace JF, et al. Development and validation of a grading system for the quality of cost-effectiveness studies. Med Care. 2003;41(1):32–44.

16. Drummond M, Manca A, Sculpher M. Increasing the generalizability of economic evaluations: recommendations for the design, analysis, and reporting of studies. Int J Technol Assess Health Care. 2005;21(2):165–71.

17. Nuckols TK, Keeler E, Morton S, et al. Economic Evaluation of Quality Improvement Interventions Designed to Prevent Hospital Readmission A Systematic Review and Meta-analysis. JAMA Intern Med. 2017 Jul;177(7):975–85.

18. Nuckols TK, Keeler E, Morton SC, et al. Economic Evaluation of Quality Improvement Interventions for Bloodstream Infections Related to Central Catheters A Systematic Review. JAMA Intern Med. 2016 Dec 1;176(12):1843–54.

19. Torre GL, Nicolotti N, Waure C de, Ricciardi W. Development of a weighted scale to assess the quality of cost-effectiveness studies and an application to the economic evaluations of tetravalent HPV vaccine. J Public Health. 2011 Apr 1;19(2):103–11.

20. Jefferson T, Demicheli V, Vale L. Quality of systematic reviews of economic evaluations in health care. JAMA. 2002;287(21):2809–12.

21. Neumann PJ, Thorat T, Shi J, Saret CJ, Cohen JT. The changing face of the cost-utility literature, 1990–2012. Value Health. 2015;18(2).

22. Siegel JE, Weinstein MC, Russell LB, Gold MR. Recommendations for reporting cost-effectiveness analyses. Panel on Cost-Effectiveness in Health and Medicine. JAMA. 1996;276(16):1339–41.

23. Russell LB, Gold MR, Siegel JE, Daniels N, Weinstein MC. The role of cost-effectiveness analysis in health and medicine. Panel on cost-effectiveness in health and medicine. JAMA. 1996;276(14):1172–7.

24. Weinstein MC, Siegel JE, Gold MR, Kamlet MS, Russell LB. Recommendations of the Panel on Cost-effectiveness in Health and Medicine. JAMA. 1996;276(15):1253–8.

25. Siler K, Lee K, Bero L. Measuring the effectiveness of scientific gatekeeping. Proc Natl Acad Sci U S A. 2015;112(2):360–5.

26. Lee CJ, Sugimoto CR, Zhang G, Cronin B. Bias in peer review. J Am Soc Inf Sci Technol. 2013;64(1):2–17.

27. Lutz Bornmann, Hans-Dieter Daniel. What do citation counts measure? A review of studies on citing behavior. J Doc. 2008;64(1):45–80.

28. Dik J-WH, Vemer P, Friedrich AW, et al. Financial evaluations of antibiotic stewardship programs—a systematic review. Front Microbiol. 2015:16(6):317-333.

29. Foster WJ, Tufail W, Issa AM. The quality of pharmacoeconomic evaluations of age-related macular degeneration therapeutics: a systematic review and quantitative appraisal of the evidence. Br J Ophthalmol. 2010;94(9):1118–26.

30. Au F, Prahardhi S, Shiell A. Reliability of two instruments for critical assessment of economic evaluations. Value Health. 2008;11(3):435–9.

31. National Institute for Health and Care Excellence. Guide to the methods of technology appraisal 2013. Available from: https://www.nice.org.uk/process/pmg9/chapter/Foreword. [Accessed July 25, 2017].

32. Drummond M, Houwing N, Slothuus U, Giangrande P. Making economic evaluations more helpful for treatment choices in haemophilia. Haemophilia. 2017 Mar;23(2):E58–66.