OPEN ACCESS Freely available online

PLoS one

External Validation of a Measurement Tool to Assess Systematic Reviews (AMSTAR)

Beverley J. Shea[1,2]*, Lex M. Bouter[3], Joan Peterson[4], Maarten Boers[5], Neil Andersson[1,6], Zulma Ortiz[7], Tim Ramsay[4], Annie Bai[8], Vijay K. Shukla[8], Jeremy M. Grimshaw[4]

1 Community Information and Epidemiological Technologies (CIET), Ottawa, Ontario, Canada, **2** Institute for Research in Extramural Medicine (EMGO Institute), Vrije Universiteit (VU) University Medical Center, Amsterdam, The Netherlands, **3** Executive Board, Vrije Universiteit (VU) University Amsterdam, Amsterdam, The Netherlands, **4** Clinical Epidemiology Program, Ottawa Health Research Institute, University of Ottawa, Ontario, Canada, **5** Department of Clinical Epidemiology and Biostatistics, Vrije Universiteit (VU) University Medical Center, Amsterdam, The Netherlands, **6** Centro de Investigación de Enfermedades Tropicales (CIET), Universidad Autónoma de Guerrero, Acapulco, Mexico, **7** Epidemiological Research Institute, National Academy of Medicine, Buenos Aires, Argentina, **8** Canadian Agency for Drugs and Technologies in Health (CADTH), Ottawa, Ontario, Canada

Background. **Thousands of systematic reviews have been conducted in all areas of health care. However, the methodological quality of these reviews is variable and should routinely be appraised. AMSTAR is a measurement tool to assess systematic reviews.** ***Methodology.*** **AMSTAR was used to appraise 42 reviews focusing on therapies to treat gastro-esophageal reflux disease, peptic ulcer disease, and other acid-related diseases. Two assessors applied the AMSTAR to each review. Two other assessors, plus a clinician and/or methodologist applied a global assessment to each review independently.** ***Conclusions.*** **The sample of 42 reviews covered a wide range of methodological quality. The overall scores on AMSTAR ranged from 0 to 10 (out of a maximum of 11) with a mean of 4.6 (95% CI: 3.7 to 5.6) and median 4.0 (range 2.0 to 6.0). The inter-observer agreement of the individual items ranged from moderate to almost perfect agreement. Nine items scored a kappa of >0.75 (95% CI: 0.55 to 0.96). The reliability of the total AMSTAR score was excellent: kappa 0.84 (95% CI: 0.67 to 1.00) and Pearson's R 0.96 (95% CI: 0.92 to 0.98). The overall scores for the global assessment ranged from 2 to 7 (out of a maximum score of 7) with a mean of 4.43 (95% CI: 3.6 to 5.3) and median 4.0 (range 2.25 to 5.75). The agreement was lower with a kappa of 0.63 (95% CI: 0.40 to 0.88). Construct validity was shown by AMSTAR convergence with the results of the global assessment: Pearson's R 0.72 (95% CI: 0.53 to 0.84). For the AMSTAR total score, the limits of agreement were -0.19 ± 1.38. This translates to a minimum detectable difference between reviews of 0.64 'AMSTAR points'. Further validation of AMSTAR is needed to assess its validity, reliability and perceived utility by appraisers and end users of reviews across a broader range of systematic reviews.**

Citation: Shea BJ, Bouter LM, Peterson J, Boers M, Andersson N, et al (2007) External Validation of a Measurement Tool to Assess Systematic Reviews (AMSTAR). PLoS ONE 2(12): e1350. doi:10.1371/journal.pone.0001350

INTRODUCTION

High quality systematic reviews are increasingly recognized as providing the best evidence to inform health care practice and policy [1]. The quality of a review, and so its worth, depends on the extent to which, scientific review methods were used to minimize the risk of error and bias. The quality of published reviews can vary considerably, even when they try to answer the same question [2]. As a result, it is necessary to appraise their quality (as is done for any research study) before the results are implemented into clinical or public health practice. Much has been written on how best to appraise systematic reviews, and while there is some variation on how this is achieved, most agree on key components of the critical appraisal [3]. Methodological quality can be defined as the extent to which the design of a systematic review will generate unbiased results [4].

Several instruments exist to assess the methodological quality of systematic reviews [5], but not all of them have been developed systematically or empirically validated and have achieved general acceptance. The authors of this paper acknowledge that the methodological quality and reporting quality for systematic reviews is very different. The first, *methodological quality*, considers how well the systematic review was conducted (literature searching, pooling of data, etc.). The second, *reporting quality*, considers how well systematic reviewers have reported their methodology and findings. Existing instruments often try to include both types of methods without being conceptually clear about the differences.

In an attempt to achieve some consistency in the evaluation of systematic reviews we have developed a tool to assess their methodological quality. This builds on previous work [6], and is based on empirical evidence and expert consensus. A measurement tool to assess systematic reviews (AMSTAR) was highly rated in a recent review (personal communication) of quality assessment instruments performed by the Canadian Agency for Drugs and Technologies in Health (CADTH). In this study we present the results of an external validation of AMSTAR using data from a series of systematic reviews obtained from the gastroenterology literature.

METHODS

The characteristics and basic properties of the instrument have been described elsewhere [7]. Briefly, a 37-item initial assessment tool was formed by combining a) the enhanced Overview Quality Assessment Questionnaire (OQAQ) scale, b) a checklist created by Sacks, and c) three additional items recently judged by experts in

Academic Editor: Joel Gagnier, University of Toronto, Canada

Received April 17, 2007; **Accepted** October 22, 2007; **Published** December 26, 2007

Funding: The authors have no support or funding to report.

Competing Interests: The authors have declared that no competing interests exist.

* **To whom correspondence should be addressed.** E-mail: bshea@ciet.org

the field to be of methodological importance. In its development phase the instrument was applied to 99 paper-based and 52 electronic systematic reviews [6] [7]. Exploratory factor analysis was used to identify underlying components. The results were considered by methodological experts using a nominal group process to reduce the number of items and design an assessment tool with face and content validity. This process lead to an 11-item instrument [7]. A description of the instrument is provided in Annex S1.

External validity

For our validation test set we chose to use systematic reviews or meta-analyses in the area of gastroenterology, specifically upper gastrointestinal. CADTH's informational specialist searched electronic bibliographic databases (i.e. Medline, Central and EMBASE) up to and including 2005. A total of 42 systematic reviews met the *a priori* criteria and were included [8]. This sample included seven electronic Cochrane systematic reviews and 35 paper-based non-Cochrane reviews. The topics of the reviews ranged across the spectrum of GI problems like dyspepsia, gastro-esophageal reflux disease (GERD), peptic ulcer disease (PUD), and also GI drug interventions such as H2 receptor antagonists and proton pump inhibitors [9–50].

Two CADTH assessors from two review groups (SS and FA, AL and CY) independently applied AMSTAR to each review and reached agreement on the assessment results. To assess construct validity, two reviewers (JP, ZO) plus a clinician and/or methodologist (MB, DF, DP, MO, and DH) applied a global assessment to each review [51] (Annex S2).

Agreement and reliability

We calculated an overall agreement score using the weighted Cohen's kappa, as well as one for each item [52] (Table 1). Bland and Altman's limits of agreement methods were used to display agreement graphically [53], [54] (Fig. 1). We calculated the percentage of the theoretical maximum score. Pearson's Rank correlation coefficients were used to assess reliability of this total score. For comparisons of rating the methodological quality we calculated chance-corrected agreement (using kappa) and chance-independent agreement (using Φ) [52], [55], [56]. We accepted a correlation of >0.66. We further scrutinized items and reviews with kappa scores below 0.66 [52]. Kappa values of less than 0 rate as less than chance agreement; 0.01–0.20 slight agreement; 0.21–0.40 fair agreement; 0.41–0.60 moderate agreement; 0.61–0.80 substantial agreement; and 0.81–0.99 almost perfect agreement [52], [57]. We calculated PHI Φ for each question [55], [58].

Construct validity

We assessed construct validity (i.e. evaluation of a hypothesis about the expected performance of an instrument) by converting the total mean score (mean of the two assessors) for each of the 42 reviews to a percentage of the maximum score for AMSTAR and of the maximum score of the global assessment instrument. We used Pearson's Rank correlation coefficients, Pearson's R and Kruskal-Wallis test to further explore the impact of the following items on the construct validity of AMSTAR: a) Cochrane systematic review vs. non-Cochrane systematic reviews [59], [60], b) journal type [61], c) year of publication [62], d) conflict of interest [63], e) impact factor [64], and number of pages [64]. We studied these in the context of a priori hypotheses concerning the correlation of AMSTAR scores. Because of the nature of their development, we anticipated that Cochrane systematic reviews would have higher quality scores than non-Cochrane systematic reviews and those electronic or general journals would score higher than specialist journals. We reported on impact factors for these journals. We hypothesized that reviews published more recently would be of higher quality than those published earlier. In addition, we anticipated that reviews declaring a conflict of interest might have lower quality scores [63], [64].

We assessed the practicability of the new instrument by recording the time it took to complete scoring and the instances where scoring was difficult. We interviewed assessors (N = 6) to obtain data on clarity, ambiguity, completeness and user-friendliness.

We used SPSS (versions 13 and 15) and MedCalc for Windows, version 8.1.0.0.

RESULTS

The 42 reviews included in the study had a wide range of quality scores. The overall scores estimated by the AMSTAR instrument ranged from 0 to 10 (out of a maximum of 11) with a mean of 4.6 (95% CI: 3.7 to 5.6; median 4.0 (range 2.0 to 6.0). The overall scores for the global assessment instrument ranged from 2 to 7 (out of a maximum score of seven) with a mean of 4.43 (95% CI: 3.6 to 5.3) and median 4.0 (range 2.5 to 5.3).

Table 1. Assessment of the inter-rater agreement for AMSTAR

Items	Kappa (95% CI)	PHI Φ
1. Was an 'a priori' design provided?	0.75 (0.55 to 0.96)	0.76
2. Was there duplicate study selection and data extraction?	0.81 (0.63 to 0.99)	0.83
3. Was a comprehensive literature search performed?	0.88 (0.73 to 1.00)	0.89
4. Was the status of publication (i.e. grey literature) used as an inclusion criterion?	0.64 (0.40 to 0.88)	0.64
5. Was a list of studies (included and excluded) provided?	0.84 (0.67 to 1.00)	0.84
6. Were the characteristics of the included studies provided?	0.76 (0.55 to 0.96)	0.76
7. Was the scientific quality of the included studies assessed and documented?	0.90 (0.77 to 1.00)	0.91
8. Was the scientific quality of the included studies used appropriately in formulating conclusions?	0.51 (0.25 to 0.78)	0.56
9. Were the methods used to combine the findings of studies appropriate?	0.80 (0.63 to 0.99)	0.80
10. Was the likelihood of publication bias assessed?	0.85 (0.64 to 1.00)	0.85
11. Were potential conflicts of interest included?	1.00 (100% no)	1.00
Overall Score	0.84 (0.67 to 1.00)	0.85

doi:10.1371/journal.pone.0001350.t001

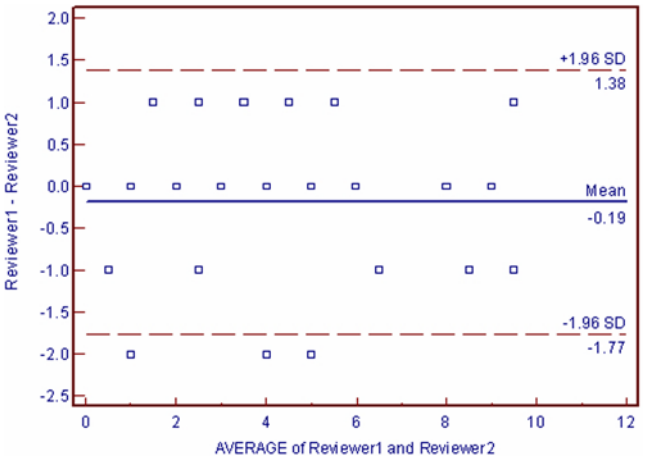

Figure 1. Bland and Altman limits of agreement plot for AMSTAR scores.
doi:10.1371/journal.pone.0001350.g001

Agreement and Reliability

The reliability of the total AMSTAR score between two assessors (the sum of all items answered 'yes' scored as 1, all others as 0) was (kappa 0.84 (95% CI: 0.67 to 1.00, $\Phi = 0.85$) and Pearson's R 0.96 (95% CI: 0.92 to 0.98). The inter-rater agreement (kappa) between two raters, for the global assessment was 0.63 (95% CI: 0.40 to 0.88).

Items in AMSTAR displayed levels of agreement that ranged from moderate to almost perfect; nine items scored a kappa of >0.75 (0.55 to 0.96 (and Φ >0.76). Item 4 had a kappa of 0.64 (0.40 to 0.88) $\Phi = 0.64$ and item 8 a kappa of 0.51(0.25 to 0.78 $\Phi = 0.56$). The reliability of the total AMSTAR score was excellent (kappa 0.84 (95% CI: 0.67 to 1.00 and Pearson's R 0.96 (95% CI: 0.92 to 0.98). For the AMSTAR total score, the limits of agreement were -0.19 ± 1.38 (Fig. 1).

The mean age of our reviewers was 40.57, median 43. Fifty-seven percent were identified as experts in methodology and 43% were identified as content experts in the field.

Construct validity

Expressed as a percentage of the maximum score, the results of AMSTAR converged with the results of the global assessment instrument [Pearson's Rank Correlation Coefficient 0.72 (95% CI: 0.53 to 0.84)]. AMSTAR scoring also upheld our other *a priori* hypotheses. The sub-analysis revealed that Cochrane reviews had significantly higher scores than paper-based reviews with a (R = 37.21 n = 7) for Cochrane reviews and (R = 18.36 n = 35) for paper-based (P<0.0002). Cochrane reviews (R = 37.21 n = 7) also scored higher than reviews published in general journals (R = 25.77 n = 11) and specialty journals (R = 14.96, n = 24) (P<0.0001). Reviews published from 2000 onward had higher AMSTAR scores than earlier reviews (R = 25.20, n = 25 vs. R = 13.12, n = 17; P = 0.0002).

The journals had the following overall summary statistics for the impact factors: mean 5.88 (95% CI: 3.9 to 7.9) median 3.3 (lowest value 1.4, highest value 23.9). There is no statistical association between AMSTAR score and impact factor (Pearson's R (0.555 P = 0.7922)). There was however a significant association found with the number of pages and AMSTAR scores (Pearson's R (0.5623 P = 0.0001 n = 42). We found no association (R 0.1773 P = 0.0308) when we removed the outliers (i.e. systematic reviews with higher page numbers).

Conflict of interest was poorly presented. Of the 42 reviews assessed, no study had appropriately declared their conflict of interest. Therefore, we were unable to assess whether or not funding had a positive or negative effect on the AMSTAR score.

Practicability

Both AMSTAR and the global assessment required on average 15 minutes to complete, but with the latter, assessors expressed difficulty in reaching a final decision in the absence of comprehensive guidelines. In contrast, AMSTAR was well received.

DISCUSSION

Principal findings

This paper describes an external validation of AMSTAR. This new measurement tool to assess methodological quality of systematic reviews showed satisfactory inter-observer agreement, reliability and construct validity in this study. Items in AMSTAR displayed levels of agreement that ranged from moderate to almost perfect. The reliability of the total AMSTAR score was excellent. Construct validity was shown by AMSTAR convergence with the results of the global assessment instrument.

We found a significant association between number of published pages and overall AMSTAR score, suggesting that the longer the manuscript, the higher the quality score. It should be interpreted with caution given the fact that only a couple of the longer reviews largely drive the hypothesis tests. We found no association when the outliers were removed from the dataset. We did not find an association between AMSTAR score and impact factor.

The AMSTAR instrument was developed pragmatically using previously published tools and expert consensus. The original 37 items were reduced to an 11- item instrument addressing key domains; the resulting instrument was judged by the expert panel to have face and content validity [7].

Strengths and weaknesses of the study

This is a prospective external validation study. We compared the new instrument to an independent and reliable gold standard designed for assessing the quality of systematic reviews, allowing multiple testing of convergent validity.

The analytical methods for assessing quality and measuring agreement amongst assessors need further discussion and development. We calculated chance-corrected agreement, using the kappa statistic [57], [65]. While avoiding high levels of agreement due to chance, kappa has its own limitations that have lead to academic criticism [66], [67]. One of the major difficulties with kappa is that when the proportion of positive ratings is extreme, the possible agreement above chance agreement is small and it is difficult to achieve even moderate values of kappa. Thus, if one uses the same raters in a variety of settings, as the proportion of positive ratings becomes extreme, kappa will decrease even if the manner in which the assessors rate the quality does not change. To address this limitation, we also calculated chance-independent agreement using PHIΦ, a relatively new approach to assessing observer agreement [55], [58].

We were unable to test our convergent validity hypothesis about conflict of interest because of missing data in the systematic reviews and primary studies. This highlights the need for journals and journal editors to require that the information is provided.

Our results are based on a small sample of systematic reviews in a particular clinical area and a relatively small number of AMSTAR assessors. There is a need for replication in larger and different data sets with more diverse appraisers.

Possible mechanisms and implications for clinicians or policymakers

Existing systematic review appraisal instrument did not reflect current evidence on potential sources of bias in systematic reviews and were generally not validated. The best available instrument prior to the development of AMSTAR was OQAQ which was formally validated. However, users of OQAQ frequently had to develop their own rules for operationalizing the instrument and OQAQ does not reflect current evidence on sources of potential bias in systematic reviews (for example funding source and conflict of interest [68,69,70]).

Quality assessment instruments can focus on either *reporting quality* (how well systematic reviewers have reported their methodology and findings (internal validity) or *methodological quality* (how well the systematic review was conducted (literature searching, pooling of data, etc.). It is possible for a systematic review with poor methodological quality to have good reporting quality. For this reason, the AMSTAR items focus on methodological quality.

Decision-makers have spent the last ten years trying to work out the best way to use the enormous amounts of systematic reviews available to them. They can hardly know where to start when deciding whether the relevant literature is valid and of the highest quality. AMSTAR is a user friendly methodological quality assessment that has the potential to standardize appraisal of systematic reviews. Early experience suggests that relevant groups are finding the instrument useful.

Unanswered questions and future research

Further validation of AMSTAR is needed to assess its validity, reliability and perceived utility by appraisers and end users of reviews across a broader range of systematic reviews. We need to assess the responsiveness of AMSTAR looking at its sensitivity to discriminate between high and low methodological quality reviews.

We need to assess the applicability of AMSTAR for reviews of observational (diagnostic, etiological and prognostic) studies and if necessary develop AMSTAR extensions for these reviews.

We plan to update AMSTAR as new evidence regarding sources of bias within systematic reviews becomes available.

SUPPORTING INFORMATION

Annex S1 AMSTAR is a measurement tool created to assess the methodological quality of systematic reviews.
Found at: doi:10.1371/journal.pone.0001350.s001 (0.04 MB DOC)

Annex S2 Global assessment rating
Found at: doi:10.1371/journal.pone.0001350.s002 (0.03 MB DOC)

ACKNOWLEDGMENTS

We would like to thank our International panel of assessors: Daniel Francis, David Henry, Marisol Betancourt, Dana Paul, Martin Olmos, and our local team of assessors: Sumeet Singh, Avtar Lal, Changhua Yu, Fida Ahmed. We also thank Dr. Giuseppe G.L. Biondi-Zoccai and Crystal Huntly-Ball for their helpful suggestions on this manuscript.

Author Contributions

Conceived and designed the experiments: JG MB BS NA LB. Performed the experiments: ZO JP VS AB. Analyzed the data: BS TR. Wrote the paper: ZO JG MB BS LB. Other: Designed the study: JS. Wrote the first draft of the paper: BS.

REFERENCES

1. Young D (2005) Policymakers, experts review evidence-based medicine. Am. J. Health Syst Pharm 62(4): 342–343.
2. Dolan-Mullen P, Ramírez G (2006) The Promise and Pitfalls of Systematic Reviews. Annual Review of Public Health 27: 81–102.
3. Oxman AD, Guyatt GH (1991) Validation of an index of the quality of review articles. J Clin Epidemiol 44(11): 1271–78.
4. Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, et al. (1995) Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. Control Clin Trials 16(1): 62–73.
5. Shea B, Dube C, Moher D (2001) Assessing the quality of reports of systematic reviews: the QUOROM statement compared to other tools. Systematic review in health care meta-analysis in context. London: BMJ Books (7): 122–39.
6. Shea B (1999) Assessing the quality of reporting meta-analyses of randomized controlled trials. MSc thesis. University of Ottawa, Department of Epidemiology and Community Medicine.
7. Shea B, Grimshaw JM, Wells GA, Boers M, Andersson N, et al. (2007) Development of AMSTAR: A Measurement Tool to Assess Systematic Reviews. BMC Medical Research Methodology 7: 10, doi:10.1186/1471-2288-7-10.
8. Singh S, Bai A, Lal A, Yu C, Ahmed F, et al. (2006) Developing evidence-based best practices for the prescribing and use of proton pump inhibitors in Canada. Ottawa, Canada: The Canadian Agency for Drugs and Technologies in Health (CADTH).
9. Chiba N, De Gara CJ, Wilkinson JM, Hunt RH (1997) Speed of healing and symptom relief in grade II to IV gastroesophageal reflux disease: a meta-analysis. Gastroenterology 112(6): 1798–810.
10. Caro JJ, Salas M, Ward A (2001) Healing and relapse rates in gastroesophageal reflux disease treated with the newer proton-pump inhibitors lansoprazole, rabeprazole, and pantoprazole compared with omeprazole, ranitidine, and placebo: evidence from randomized clinical trials. Clin Ther 23(7): 998–1017.
11. Klok RM, Postma MJ, van Hout BA, Brouwers JR (2003) Meta-analysis: comparing the efficacy of proton pump inhibitors in short-term use. Aliment Pharmacol Ther 17(10): 1237–45.
12. Van Pinxteren B, Numans ME, Lau J, de Wit NJ, Hungin AP, et al. (2003) Short-term treatment of gastroesophageal reflux disease. J Gen Intern Med 18(9): 755–63.
13. Van Pinxteren B, Numans ME, Bonis PA, Lau J (2004) Short-term treatment with proton pump inhibitors, H2-receptor antagonists and prokinetics for gastro-oesophageal reflux disease-like symptoms and endoscopy negative reflux disease. Cochrane Database Syst Rev (3): CD002095.
14. Rostom A, Dube C, Wells G, Tugwell P, Welch V, et al. (2002) Prevention of NSAID-induced gastroduodenal ulcers. Cochrane Database Syst Rev (4): CD002296.
15. Laheij RJ, van Rossum LG, Jansen JB, Straatman H, Verbeek AL (1999) Evaluation of treatment regimens to cure Helicobacter pylori infection: a meta-analysis. Aliment Pharmacol Ther 13(7): 857–64.
16. Carlsson R, Galmiche JP, Dent J, Lundell L, Frison L (1997) Prognostic factors influencing relapse of oesophagitis during maintenance therapy with antisecretory drugs: a meta-analysis of long-term omeprazole trials. Aliment Pharmacol Ther 11(3): 473–82.
17. Chiba N (1997) Proton pump inhibitors in acute healing and maintenance of erosive or worse esophagitis: a systematic overview. Can J Gastroenterol 11 Suppl B: 66B–73B.
18. Delaney B, Moayyedi P, Deeks J, Innes M, Soo S, et al. (2000) The management of dyspepsia: a systematic review. Health Technol Assess 4(39); i,iii-189. Available: http://www.ncchta.org/execsumm/summ439.htm.
19. Moayyedi P, Soo S, Deeks J, Delaney B, Harris A, et al. (2005) Eradication of Helicobacter pylori for non-ulcer dyspepsia. Cochrane Database Syst Rev (1): CD002096.
20. Moayyedi P, Soo S, Deeks J, Delaney B, Innes M, et al. (2005) Pharmacological interventions for non-ulcer dyspepsia. Cochrane Database Syst Rev (1): CD001960. Available: http://www.mrw.interscience.wiley.com/cochrane/clsysrev/articles/CD001960/pdf_fs.html (accessed 2006 Feb 16).
21. Delaney BC, Moayyedi P, Forman D (2003) Initial management strategies for dyspepsia. Cochrane Database Syst Rev (2): CD001961.
22. Hopkins RJ, Girardi LS, Turney EA (1996) Relationship between Helicobacter pylori eradication and reduced duodenal and gastric ulcer recurrence: a review. Gastroenterology 110(4): 1244–52.
23. Huang JQ, Sridhar S, Hunt RH (2002) Role of Helicobacter pylori infection and non-steroidal anti-inflammatory drugs in peptic-ulcer disease: a meta-analysis. Lancet 359(9300): 14–22.

24. Moayyedi P, Soo S, Deeks J, Forman D, Mason J, et al. (2000) Systematic review and economic evaluation of Helicobacter pylori eradication treatment for non-ulcer dyspepsia. Dyspepsia Review Group. BMJ 321(7262): 659–64.
25. Jovell AJ, Aymerich M, Garcia Altes A, Serra Prat M (1998) Clinical practice guideline for the eradicating therapy of Helicobacter pylori infections associated to duodenal ulcer in primary care. Barcelona: Catalan Agency for Health Technology Assessment. Available: http://www.gencat.net/salut/depsan/units/aatrm/pdf/gp9802en.pdf.
26. Gisbert JP, González L, Calvet X, García N, López T (2000) Proton pump inhibitor, clarithromycin and either amoxycillin or nitroimidazole: a meta-analysis of eradication of Helicobacter pylori. Aliment Pharmacol Ther 14(10): 1319–28.
27. Calvet X, García N, López T, Gisbert JP, Gené E, et al. (2000) A meta-analysis of short versus long therapy with a proton pump inhibitor, clarithromycin and either metronidazole or amoxycillin for treating Helicobacter pylori infection. Aliment Pharmacol Ther 14(5): 603–09.
28. Gené E, Calvet X, Azagra R, Gisbert JP (2003) Triple vs. quadruple therapy for treating Helicobacter pylori infection: a meta-analysis. Aliment Pharmacol Ther 17(9): 1137–43.
29. Huang J, Hunt RH (1999) The importance of clarithromycin dose in the management of Helicobacter pylori infection: a meta-analysis of triple therapies with a proton pump inhibitor, clarithromycin and amoxycillin or metronidazole. Aliment Pharmacol Ther 13(6): 719–29.
30. Leodolter A, Kulig M, Brasch H, Meyer Sabellek W, Willich SN, et al. (2001) A meta-analysis comparing eradication, healing and relapse rates in patients with Helicobacter pylori-associated gastric or duodenal ulcer. Aliment Pharmacol Ther 15(12): 1949–58.
31. Moayyedi P, Murphy B (2001) Helicobacter pylori: a clinical update. J Appl Microbiol (30): 126S–33S.
32. Oderda G, Rapa A, Bona G (2000) A systematic review of Helicobacter pylori eradication treatment schedules in children. Aliment Pharmacol Ther 14(Suppl 3): 59–66.
33. Schmid CH, Whiting G, Cory D, Ross SD, Chalmers TC (1999) Omeprazole plus antibiotics in the eradication of Helicobacter pylori infection: a meta-regression analysis of randomized, controlled trials. Am J Ther 6(1): 25–36.
34. Unge P, Berstad A (1996) Pooled analysis of anti-Helicobacter pylori treatment regimens. Scand J Gastroenterol Suppl 220: 27–40.
35. Unge P (1998) Antimicrobial treatment of H. pylori infection: a pooled efficacy analysis of eradication therapies. Eur J Surg Suppl 582: 16–26.
36. Unge P (1997) What other regimens are under investigation to treat Helicobacter pylori infection? Gastroenterology 113(6 Suppl): S131–S148.
37. Vallve M, Vergara M, Gisbert JP, Calvet X (2002) Single vs. double dose of a proton pump inhibitor in triple therapy for Helicobacter pylori eradication: a meta-analysis. Aliment Pharmacol Ther 16(6): 1149–56.
38. Veldhuyzen van Zanten SJ, Sherman PM (1994) Indications for treatment of Helicobacter pylori infection: a systematic overview. CMAJ 150(2): 189–98.
39. Trépanier EF, Agro K, Holbrook AM, Blackhouse G, Goeree R, et al. (1998) Meta-analysis of H pylori (HP) eradication rates in patients with duodenal ulcer (DU). Can J Clin Pharmacol. 5(1): 67.
40. Bamberg P, Caswell CM, Frame MH, Lam SK, Wong EC (1992) A meta-analysis comparing the efficacy of omeprazole with H2-receptor antagonists for acute treatment of duodenal ulcer in Asian patients. J Gastroenterol Hepatol 7(6): 577–85.
41. Di Mario F, Battaglia G, Leandro G, Grasso G, Vianello F, et al. (1996) Short-term treatment of gastric ulcer: a meta-analytical evaluation of blind trials. Dig Dis Sci 41(6): 1108–31.
42. Eriksson S, Langstrom G, Rikner L, Carlsson R, Naesdal J (1995) Omeprazole and H2-receptor antagonists in the acute treatment of duodenal ulcer, gastric ulcer and reflux oesophagitis: a meta-analysis. Eur J Gastroenterol Hepatol 7(5): 467–75.
43. Poynard T, Lemaire M, Agostini H (1995) Meta-analysis of randomized clinical trials comparing lansoprazole with ranitidine or famotidine in the treatment of acute duodenal ulcer. Eur J Gastroenterol Hepatol 7(7): 661–65.
44. Laine L, Schoenfeld P, Fennerty MB (2001) Therapy for Helicobacter pylori in patients with nonulcer dyspepsia: a meta-analysis of randomized, controlled trials. Ann Intern Med 134(5): 361–9.
45. Mulder CJ, Schipper DL (1990) Omeprazole and ranitidine in duodenal ulcer healing. Analysis of comparative clinical trials. Scand J Gastroenterol Suppl 178: 62–6.
46. Shiau JY, Shukla VK, Dubé C (2002) The efficacy of proton pump inhibitors in adults with functional dyspepsia. Ottawa: Canadian Coordinating Office for Health Technology Assessment.
47. Danesh J, Lawrence M, Murphy M, Roberts S, Collins R (2005) Systematic review of the epidemiological evidence on Helicobacter pylori infection and non-ulcer or uninvestigated dyspepsia. Arch Intern Med 160(8): 1192–98.
48. Gibson PG, Henry RL, Coughlan JL (2005) Gastro-esophageal reflux treatment for asthma in adults and children. Cochrane Database Syst Rev (3): 1–27.
49. Fischbach LA, Goodman KJ, Feldman M, Aragaki C (2002) Sources of variation of helicobacter pylori treatment success in adults worldwide: a meta-analysis. Int J Epidemiol 31(1): 128–39.
50. Ford A, Delaney B, Moayyedi P (2003) Eradication therapy for peptic ulcer disease in helicobacter pylori positive patients. Cochrane Database Syst Rev (4): CD003840.
51. Oxman AD, Guyatt GH (1991) Validation of an index of the quality of review articles. J Clin Epidemiol 44(11): 1271–78.
52. Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Measure 0(1): 622–26.
53. Bland JM, Altman DG (1986) Statistical methods for assessing agreement between two methods of clinical Measurement. Lancet i: 307–10.
54. Bland JM, Altman DG (1987) Statistical methods for assessing agreement between measurement. Biochimica Clinica 11: 399–404.
55. Meade M, Cook R, Guyatt G, Groll R, Kachura J, et al. (2000) Interobserver Variation in Interpreting Chest Radiographs for the Diagnosis of Acute Respiratory Distress Syndrome. Am. J. Respir. Crit. Care Med 161(1): 185–90.
56. Uebersax JS (1987) Diversity of decision-making models and the measurement of inter-rater agreement. Psychological Bulletin 101: 140–46.
57. Cohen J (1968) Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 70: 213–220.
58. McGinn T, Guyatt G, Cook R, Meade M (2002) Diagnosis: measuring agreement beyond chance. In: Guyatt G, Rennie D, eds. Users' guide to the medical literature. A manual for evidence-based clinical practice. Chicago, IL: AMA Press. pp 461–70.
59. Moja LP, Telaro E, D'Amico R, Moschetti I, Coe L, et al. (2005) Assessment of methodological quality of primary studies by systematic reviews: results of the metaquality cross sectional study. BMJ Publishing Group Ltd. 330(7499): 1053.
60. Shea B, Moher D, Graham I, Pham B, Tugwell P (2002) A comparison of the quality of Cochrane reviews and systematic reviews published in paper-based journals. Evaluation & the Health Professions 25(1): 116–29.
61. Moher D, Tetzlaff J, Tricco AC, Sampson M, Altman DG (2007) Epidemiology and Reporting Characteristics of Systematic Reviews. PLoS Med 4(3): e78, doi:10.1371/journal.pmed.0040078.
62. Sacks HS, Berrier J, Reitman D, Ancona-Berk VA, Chalmers TC (1987) Meta-analyses of randomized controlled trials. New England Journal of Medicine 316: 450–54.
63. Bero LA (2005) Managing financial conflicts of interest in research. Journal of the American College of Dentists 72(2): 4–9.
64. Biondi-Zoccai G, Lotrionte M, Abbate A, Testa L (2006) Compliance with QUOROM and quality of reporting of overlapping meta-analyses on the role of acetylcysteine in the prevention of contrast associated nephropathy: case study. BMJ 332: 202–6.
65. Fleiss JL (1971) Measuring nominal scale agreement among many raters. Psychol. Bull 76: 378–382.
66. McClure M, Willett W (1987) Misinterpretation and misuse of the kappa statistic. Am. J. Epidemiol. 126: 161–169.
67. Cook RJ, Farewell VT (1995) Conditional inference for subject-specific and marginal agreement: two families of agreement measures. Can. J. Stat 23: 333–344.
68. Barnes DE, Bero LA (1998) Why review articles on the health effects of passive smoking reach different conclusions. JAMA 279: 1566–1570.
69. Cho MK, Bero LA (1996) The quality of drug studies published in symposium proceedings. Ann Intern Med 124: 485–489.
70. Lexchin J, Bero LA, Djulbegovic B, Clark O (2003) Pharmaceutical industry sponsorship and research outcome and quality: systematic review. BMJ 326: 1167–1170.