From Systematic Reviews to Clinical Recommendations for Evidence-Based Health Care: Validation of Revised Assessment of Multiple Systematic Reviews (R-AMSTAR) for Grading of Clinical Relevance

Jason Kung [1,2], Francesco Chiappelli[*,1], Olivia O. Cajulis[3], Raisa Avezova[#,1], George Kossan[#,1], Laura Chew[#,1] and Carl A. Maida[1,4]

[1]*Division of Oral Biology & Medicine, UCLA School of Dentistry, Los Angeles CA*

[2]*Department of Dentistry, Saint Barnabas Hospital, New York, NY*

[3]*Dental Group of Sherman Oaks, Los Angeles, CA*

[4]*Division of Public Health and Community Dentistry, UCLA School of Dentistry*

Abstract: Research synthesis seeks to gather, examine and evaluate systematically research reports that converge toward answering a carefully crafted research question, which states the problem patient population, the intervention under consideration, and the clinical outcome of interest. The product of the process of systematically reviewing the research literature pertinent to the research question thusly stated is the "systematic review".

The objective and transparent approach of the systematic review aims to minimize bias. Most systematic reviews yield quantitative analyses of measurable data (e.g., acceptable sampling analysis, meta-analysis). Systematic reviews may also be qualitative, while adhering to accepted standards for gathering, evaluating, and reporting evidence. Systematic reviews provide highly rated recommendations for evidence-based health care; but, systematic reviews are not equally reliable and successful in minimizing bias.

Several instruments are available to evaluate the quality of systematic reviews. The 'assessment of multiple systematic reviews' (AMSTAR) was derived from factor analysis of the most relevant items among them. AMSTAR consists of eleven items with good face and content validity for measuring the methodological quality of systematic reviews, has been widely accepted and utilized, and has gained in reliability, reproducibility. AMSTAR does not produce quantifiable assessments of systematic review quality and clinical relevance.

In this study, we have revised the AMSTAR instrument, detracting nothing from its content and construct validity, and utilizing the very criteria employed in the development of the original tool, with the aim of yielding an instrument that can quantify the quality of systematic reviews. We present validation data of the revised AMSTAR (R-AMSTAR), and discuss its implications and application in evidence-based health care.

INTRODUCTION

The new science of research synthesis has emerged from the seminal work of Archibald Cochrane (1909-1988) in health care over the last four decades [1]. The principal aim of research synthesis is to gather, to examine and to evaluate systematically research reports that converge toward answering a carefully developed research question. The question is crafted so as to state clearly and unequivocally the problem patient population (P), the intervention (I) under consideration (C), and the clinical outcome (O) of interest. The product of the process of systematically reviewing the research literature pertinent to the research question (P.I.C.O.) thusly stated has been termed the "systematic review" [2-4].

The systematic review is not identical to the literature review because it rests upon an objective and transparent approach, which is grounded on the science of research synthesis with the specific intent and goal to minimize bias. To that end, most systematic reviews are based on an explicit quantitative analysis of measurable data (e.g., acceptable sampling analysis, meta-analysis). Despite these concerted efforts, certain threats of bias subsist (e.g., publication bias). Moreover, a substantial number of systematic reviews are qualitative in nature, and, while adhering to accepted standards for gathering, evaluating, and reporting evidence, do not yield quantitative assessments [4].

Consequently, while systematic reviews provide, in principle, the most highly rated recommendations for evidence-based health care, it is now evident that, in actuality, not all systematic reviews are equally as reliable and of sufficient quality to minimize bias satisfactorily [5]. Other sources of divergence, or "discordance", in systematic reviews arise from the fundamental need of regular updates. A brief survey of the pertinent body literature, or "bibliome", suggests that

*Address correspondence to this author at the UCLA School of Dentistry, CHS 63-090 Los Angeles, CA 90095-1668, USA: Tel: 310-794-6625; Fax: 310-794-7109; E-mail: fchiappelli@dentistry.ucla.edu

Ms. Avezova, Ms. Chew and Mr. Kossan are pre-dental students at UCLA

1874-2106/10

at least 10% of all systematic reviews in health care need updating at the time of publication because of the length of time taken in preparing a systematic review, and of the accelerated pace of scientific production of new evidence [6]. Moreover, systematic reviews may employ and incorporate "gray literature" (e.g., unpublished observations, dissertations, conference proceedings) to different extents in an effort to be all-inclusive of the available literature [13,4,7]. These disparities in protocol contribute to yielding a confused and often conflicting bibliome of discordant systematic reviews, which seems to complicate, rather than to ease the process of clinical evidence-based decision-making in health care [3,8,9].

To confront this important problem, several instruments have been developed and validated in order to evaluate the quality of systematic reviews, starting with a simple checklist over 15 years ago [10], and soon evolving into the more cumbersome Overview Quality Assessment Questionnaire (OQAQ) [11]. In an effort to develop an instrument to assess the methodological quality of systematic reviews, building upon previous tools, empirical evidence and expert consensus, a 37-item assessment tool was devised by combining the items of the OQAQ, an expanded checklist [12], and three additional items considered methodologically relevant. This hybrid tool was validated with close to 150 systematic reviews, and factor analyzed so as to identify underlying components. Items were reduced to eleven in the creation of the 'assessment of multiple systematic reviews' (AMSTAR) [13]. Careful psychometric assessments determined that AMSTAR has good face and content validity for measuring the methodological quality of systematic reviews and clinical relevance. As of this date, AMSTAR has been widely accepted and utilized by professional health care associations and other policy institutions, and has gained in respectability, reliability, reproducibility. AMSTAR, while appropriate and user-friendly, fails to produce quantifiable assessments of systematic review quality [1,8,9,13].

We have revised AMSTAR, detracting nothing from its content and construct validity, utilizing the criteria originally employed in its development, and produced the revised AMSTAR (R-AMSTAR) that successfully quantifies the quality of systematic reviews (Appendix 1). Here, we present and discuss its validation in evidence-based clinical decision-making health care.

METHOD

We selected at random two independent domains of clinical work in health care: post-traumatic stress syndrome (PTSD), and rheumatoid arthritis (RA). We appropriately crafted a P.I.C.O. question for each domains, which suggested medical subject heading keywords for inclusion and exclusion purposes. We searched the National Library or Medicine (Pubmed, Medline), Google Scholar, Cochrane Library, Center for Reviews & Dissemination, and EMBASE, and supplemented the yield by manual search of the literature. We excluded systematic reviews that were not in English, "gray literature", literature reviews that were not systematic reviews by the criteria of be centered around a clearly stated P.I.C.O. question, and any primary research report (i.e., clinical trial, observational study).

We trained four independent readers, and ensured their ability to read critically and following the criteria of R-AMSTAR (cf., Appendix 1) in a standard manner by running blind mock critical R-AMSTAR assessment sessions of the same systematic review, and comparing the outcomes. Any divergent response was discussed until consensus was reached as to how, specifically, the criteria of R-AMSTAR ought to be applied for each of the eleven domains.

Two readers focused on the PTSD systematic reviews, and the other two readers concerned themselves with the RA systematic reviews. Readings were independent and blind from each other. Data were pooled for each bibliome, averaged, and analyzed quantitatively by a fifth member of the research team to ensure unbiased and blind analyses and interpretations.

Acceptable sampling analysis was performed as described (3,8), using the Friedman test for non-parametric analysis of factorial designs, followed by non-parametric post-hoc comparisons, and Bonferroni correction of the level of significance (α=0.05) as needed (MDAS statistical software: Medical Data Analysis System, EsKay Software, Pittsburgh, 2004). In brief, scores from both readers in each bibliome were tabulated across the eleven domains of the original AMSTAR scored based on the original criteria as described in Appendix 1. Marginal totals were utilized to establish the level of acceptability within a 95% confidence interval (CI95). The relative strength of each domain was described by their respective means and coefficient of variation, and inclusion within the respective CI95 for each domain, and compared when needed, by Wilcoxon (MDAS), with Bonferroni correction as noted.

RESULTS

In systematic reviews, a flowchart is often presented that lists the process by which the bibliome of pertinent literature is progressively obtained by inclusion and exclusion criteria. Lest the present report be misconstrued as a systematic review, we do not present the information in that format here. The present study used, as its unit of research, not individual primary research reports (e.g., clinical trials), but existing systematic reviews. Whereas systematic reviews seek to identify the best available evidence within a given bibliome for or against a certain treatment intervention, the intent of the data we present here is not that: rather, it seeks to utilize a coherent and homogeneous bibliome in a health science topic to verify the validity of our approach to quantify AMSTAR-derived assessments of systematic review quality.

With this intent, suffice to say that, from an original search of systematic reviews, we obtained 394 pertinent entries for PTSD, and 970 entries for RA. Upon, filtering for intervention and for inflammation respectively, the number of PTSD and RA systematic reviews decreased respectively to 72 and 71. Further filtering (psychological treatment, inflammation bone) lowered the number of coherent systematic reviews to 20 for PTSD and 10 for RA. Upon verification of outcome homogeneity and language, a final total of 11 systematic reviews for PTSD and of 5 systematic reviews for RA were obtained, and used in the validation study.

The data shown in Table **1** list the average scores for R-AMSTAR for the PTSD literature. The inter-rate reliability

Table 1. Average R-AMSTAR Scores Across 2 Independent Readers for the PTSD Literature

Report	1	2	3	4	5	6	7	8	9	10	11	Total
1	4.00	3.50	3.50	3.00	1.50	4.00	3.50	4.00	4.00	1.00	1.50	**32.50**
2	4.00	4.00	3.50	2.50	1.50	3.00	2.50	4.00	4.00	2.50	1.00	**33.00**
3	4.00	4.00	4.00	2.00	1.00	2.50	3.00	4.00	4.00	1.00	1.50	**32.50**
4	4.00	2.50	4.00	4.00	4.00	4.00	4.00	4.00	3.50	4.00	3.50	**39.00**
5	4.00	2.50	4.00	2.00	1.50	4.00	3.50	4.00	3.00	1.50	2.00	**32.00**
6	4.00	1.00	3.50	2.00	2.00	2.50	3.00	2.50	2.00	1.50	1.50	**25.50**
7	4.00	4.00	3.50	1.50	1.00	4.00	3.00	4.00	4.00	1.50	2.50	**33.00**
8	4.00	1.50	3.50	1.50	1.50	4.00	2.50	4.00	2.50	1.50	3.00	**29.50**
9	4.00	2.00	3.00	1.00	2.00	3.50	4.00	4.00	3.00	1.00	2.50	**30.00**
10	4.00	1.00	3.50	2.00	1.00	4.00	3.00	4.00	3.00	1.50	2.00	**29.00**
11	4.00	1.00	3.50	2.50	2.50	4.00	2.50	3.50	2.00	1.50	3.00	**30.00**
Mean	**4.00**	**2.45**	**3.59**	**2.18**	**1.77**	**3.59**	**3.14**	**3.82**	**3.18**	**1.68**	**2.18**	31.45
SD	**0.00**	**1.25**	**0.30**	**0.81**	**0.88**	**0.63**	**0.55**	**0.46**	**0.78**	**0.87**	**0.78**	3.37

($p<0.0001$, Friedman non-parametric ANOVA equivalent)

Table 2. Average R-AMSTAR Scores Across 2 Independent Readers for the RA Literature

Reports	1	2	3	4	5	6	7	8	9	10	11	Total
1	4.00	1.00	4.00	2.00	3.50	3.50	3.50	3.50	1.00	1.00	1.00	**28.00**
2	3.50	2.50	4.00	4.00	3.50	4.00	3.50	2.50	3.50	1.50	1.00	**33.50**
3	4.00	4.00	3.50	4.00	1.50	2.50	3.50	3.50	2.50	1.50	1.00	**31.50**
4	4.00	2.00	4.00	4.00	2.00	4.00	3.50	3.00	3.50	1.00	1.00	**32.00**
5	3.50	4.00	4.00	3.00	2.50	4.00	4.00	4.00	2.50	1.00	2.50	**35.00**
Mean	**3.80**	**2.70**	**3.90**	**3.40**	**2.60**	**3.60**	**3.60**	**3.30**	**2.60**	**1.20**	**1.30**	**32.00**
SD	**0.27**	**1.30**	**0.22**	**0.89**	**0.89**	**0.65**	**0.22**	**0.57**	**1.02**	**0.27**	**0.67**	**2.62**

($p=0.001$, Friedman non-parametric ANOVA equivalent)

for this set was 0.58. Table **2** lists the average scores for R-AMSTAR for the RA literature, where the inter-rate reliability obtained 64% of shared variance (Pearson r=0.80).

The data in Table **1** show that all of the systematic reviews examined in PTSD had a R-AMSTAR a score that fell within the confidence interval set by the sample (mean$\pm$standard deviation: 31.45$\pm$3.37, CI95: 24.8 – 38.06), except for Report 4 (score = 39.0), and Report 6, possibly bordering the lower confidence limit (score = 25.50). As indicated in Table **3**, paper 4 ranked with highest score ("A" quality systematic review – most trustworthy consensus statement based on the best available evidence) ; papers 1-3, 5 & 7 ranked within the top 80th percentile (B quality systematic review). Papers 8-11 ranked in the 70th percentile based of the aggregate R-AMSTAR scores ("C" quality systematic review), and paper 6, in this example, presents a systematic review so flawed, based on AMSTAR criteria and R-AMSTAR quantification, that it hardly offers noteworthy clinical relevance.

Table **1** also shows a significant difference in the relative scores for each of the eleven domains of the R-AMSTAR (Friedman, $p<0.0001$). Whereas none of the R-AMSTAR questions across the PTSD bibliome showed an overall mean score outside the 95% confidence limits, domains represented by question 5 (appropriate inclusion and exclusion of the literature) and 10 (publication bias) inspire caution and limited confidence. Domains represented by questions 4 (gray literature) and 11 (conflict of interest) also appear relatively weak. Taken together, the average scores of these four domains are significantly lower than the remaining stronger domains represented by questions 1,2,3,6,7,8 & 9 (Wilcoxon, $p=0.0002$).

Table 3. Systematic Review Ranking Based on R-AMSTAR Scores

PICO	Paper	R-AMSTAR[a]	%ile	Rank[b]
PTSD	1	32.50	83	B
	2	33.00	85	B
	3	32.50	83	B
	4	39.00	100	A
	5	32.00	82	B
	6	25.50	65	D
	7	33.00	85	B
	8	29.50	76	C
	9	30.00	77	C
	5	35.00	100	A
	11	30.00	77	C
RA	1	28.00	80	B
	2	33.50	96	A
	3	31.50	90	A
	4	32.00	91	A

[a]The values listed in the table correspond to the total R-AMSTAR scores listed in Tables **1** & **2** respectively for the PTSD and the RA bibliome

[b]Based on the criteria of excellence of systematic reviews that resulted in the 11 domains examined by the AMSTAR, the overall score on the R-AMSTAR, which is revised only to the extent that it produces a quantification of the assessments of these domains, reveals the possibility to assign a grade of systematic review quality and clinical relevance, based on the criteria of the top percentile of the scores reflecting an A paper, and so on. The rankings are, for obvious reasons, relative strictly to the systematic reviews examined in response to the specific P.I.C.O. question, and thus pertain to a fixed (rather than random) interpretative model.

A similar analysis is presented in Table **2** for the RA bibliome, which presents with an overall R-AMSTAR score CI95 (26.86 - 37.14). In contrast to the PTSD bibliome, most of the systematic reviews in the RA bibliome obtained an "A" quality of systematic review, ranking at or within the 90th percentile. Paper 1 ranked as a "B" quality systematic review (80th percentile) (Table **3**).

Table **2** also shows a statistically significant difference in the relative scores among the overall scores across the eleven domains tested by R-AMSTAR across the RA bibliome (Friedman, p=0.0001). The data evince particular weaknesses in the field *vis à vis* questions 10 and 11 (gray literature inclusion and conflict of interest, respectively). The scores of these questions are significantly lower than the scores on questions 1-9 (Wilcoxon, p<0.00001).

DISCUSSION

The findings, presented in Tables **1-3**, confirm the usefulness of acceptable sampling analysis in the context of systematic reviews. Quality of the evidence assessments, when performed by a well-constructed and standardized instrument that captures the widely accepted domains of the scientific process, can generate important observations with respect to general acceptability of reports (i.e., Tables **1** & **2** vertical marginals: total scores), as well as inherent strengths and weakness of the *corpus* of research under examination. The latter are rendered by the horizontal marginal means in Tables **1** & **2**. Table **3** presents a transformation of the information in the preceding tables in such a manner as to proffer a succinct and easy-to-interpret grading system (A-D), based on the fundamental and commonly accepted concept of percentile, which clinicians and third-party providers can use in order to evaluate at a glance the evidence synthesized in he systematic review under consideration.

The process of systematic review quality assessment, evaluation and dissemination we propose here relies on the construct, content and criterion validity of the AMSTAR instrument, which has been established, described and documented by others over the past decade (10-13). The AMSTAR is commonly utilized by investigators in the health care fields, as well as policy-making associations (e.g., American Dental Association, Evidence-Based Dentistry Center).

Through factor analysis and psychometric characterization, eleven domains were obtained, which constitute the AMSTAR. These domains are commonly accepted among researchers and clinicians in evidence-based health care to assess adequately the research synthesis stringency and the clinical relevance of any given systematic review. The major flaw of the AMSTAR instrument, however, is that it generates a qualitative evaluation, and fails to quantify the systematic review quality.

In order to address this limitation, we utilized the criteria that are imbedded within each of the eleven domains of the original AMSTAR, and produced scores based on whether critical reading revealed satisfactory vs. unsatisfactory coverage of each criterion. Following a series of pilot studies aimed at refining and adjusting the relative weight of the criteria within each domain, we obtained the R-AMSTAR (cf., Appendix 1), which preserves the construct, content and criterion validity of the original instrument, while permitting quantification.

The quantified measures of the R-AMSTAR are scores on each of the individual eleven domains of the original instrument, based on the criteria discussed above. Each domain's score ranges between 1 and 4 (maximum), and the R-AMSTAR total scores has a range of 11 to 44 (maximum). By implication, a total score of 11 (e.g., Tables **1** & **2**, horizontal marginals) signifies that none of the AMSTAR criteria were satisfied along said established eleven domains. By contrast, a score of 44 reveals that all of the criteria of systematic review excellence were verified in every domain. That is to say, low R-AMSTAR total scores should lead to prudence on the part of the clinician, whereas high R-AMSTAR total scores should impart a certain degree of confidence about the clinical relevance and implications of the findings discussed in the high scoring systematic review.

Relative to the set of systematic reviews within a given bibliome (e.g., here P.I.C.O. question on PTSD, and P.I.C.O. question on RA), a ranking of systematic review quality can be obtained, and graded based on the widely accepted, simple concept of percentiles: in that manner, most of the systematic reviews responding to the RA P.I.C.O. question were deserving of an "A" grade. "A" grade systematic reviews are those that adhere stringently to commonly shared principles of clinical relevance; therefore, "A" grade systematic reviews are those, which clinicians can use with the greatest degree of confidence in the process of making evidence-based clinical decisions. It is of note that in both example bibliomes examined here, the top-ranking systematic reviews were not, as one might expect uniformally, Cochrane reviews.

The methodological strength of the R-AMSTAR further rests on the fact that an acceptable sampling analysis may be conducted both along the vertical marginal totals (i.e., total scores), which, by the adoption of some conventional criterion cut-off point (e.g., total score of 22, which indicates that, on average, only two criteria for each of the domains tested were satisfied), permits the exclusion of low scoring systematic reviews. This process, which might not be recommended in all cases, such as, for example, when a Bayesian interpretation of the best available evidence is sought, is useful in specific cases of the construction of complex systematic reviews (aka, meta-systematic reviews, 3,8).

Furthermore, analysis of the marginal means and standard deviations (horizontal marginal values in Tables **2** & **3**), yields valuable information with respect to the relative strength or weakness of a bibliome along each given domain (Tables **1** & **2**). When the reports within a bibliome are uniformly strong (or weak) along the eleven AMSTAR domains, then the Friedman analysis of the tabulated scores reveals no statistical significance. A significant Friedman analysis, such as those evinced in both Tables **1** & **2**, indicates that certain domains are strong and acceptable, while others are alarmingly weak and can seriously jeopardize the clinical relevance of the systematic reviews that overall constitute the bibliome under study. Further post-hoc analysis proffers the ability to identify these weaknesses, such that the alerted clinician, can, if so desiring, use in the clinical decision-making process even "B" or "C" systematic reviews, so long as the recommendations are interpreted in light of the identified limitations through the acceptable sampling analysis protocol just described.

Taken together, the analyses we present here improve and expand the use of the commonly accepted AMSTAR instrument (11,13) by enabling reliable quantification of the eleven domains taken to represent clinical relevance of systematic reviews. Our approach (9) permits a detailed analysis of acceptable vs. deficient aspects of each systematic review within any given bibliome obtained in answering a specific P.I.C.O. question, as well as the overall strengths and limitations of the bibliome as a *corpus* of literature. Furthermore, the R-AMSTAR we describe here (Appendix 1) yields a total score, which proffers the ability to rank and grade each systematic review relative to each other in the bibliome under study (Table **3**). The clinical utility of the A-through-D grades thus obtained, which are easy to grasp for use in evidence-based clinical decision-making, and based on the simple principle of percentiles, rests on the fact that their interpretation is reflective of the clinical relevance of any systematic review within the bibliome.

It is also the case that, in absolute terms, the overall total score of the R-AMSTAR reflects the adherence of any systematic review under evaluation to the generally accepted criteria of quality of the synthesized evidence, regardless of the P.I.C.O. question. That is to say, it is possible and even probable that the translation of the scores into percentile, as we propose in Table **3** relative to individual specific P.I.C.O. questions, can be generalized. A random (rather than fixed) interpretative model of the R-AMSTAR quantifications will bring much needed cohesion and uniformity to the field of evidence-based decision-making. In conclusion, the R-AMSTAR yields numbers based upon generally agreed upon criteria of excellence, which can be transformed into a standardized grading system of the quality of the evidence presented in any given systematic review (cf., Table **3**); and, such transformation can be generalized across the entire bibliome of research synthesis in clinical dentistry to yield a simple, reliable, and easy-to-grasp quantification of the quality of the evidence for any systematic review under consideration.

ACKNOWLEDGMENTS

The authors thank the colleagues, graduate students and undergraduate pre-dental students who have actively participated in the elaboration of the theoretical and practical construct of evidence-based research and evidence-based practice elaborated in our research team over the years. This study received no intramural or extramural funding, and the authors declare no conflicts of interest.

APPENDIX 1

Revised Amstar

1. Was an 'a priori' design provided?
If it satisfies 3 of the criteria →4
If it satisfies 2 of the criteria →3
If it satisfies 1 of the criteria →2
If it satisfies 0 of the criteria →1

Criteria:

(A)	'a priori' design
(B)	statement of inclusion criteria
(C)	PICO/PIPO research question (population, intervention, comparison, prediction, outcome)

2. Was there duplicate study selection and data extraction?

If it satisfies 3 of the criteria →4
If it satisfies 2 of the criteria →3
If it satisfies 1 of the criteria →2
If it satisfies 0 of the criteria →1

Criteria:

(A)	There should be at least two independent data extractors as stated or implied.
(B)	Statement of recognition or awareness of consensus procedure for disagreements.
(C)	Disagreements among extractors resolved properly as stated or implied

3. Was a comprehensive literature search performed?

If it satisfies 4 or 5 of the criteria → 4
If it satisfies 3 of the criteria → 3
If it satisfies 2 of the criteria →2
If it satisfies 1 or 0 of the criteria → 17

Criteria:

(A)	At least two electronic sources should be searched.
(B)	The report must include years and databases used (e.g. Central, EMBASE, and MEDLINE).
(C)	Key words and/or MESH terms must be stated **AND** where feasible the search strategy outline should be provided such that one can trace the filtering process of the included articles.
(D)	In addition to the electronic databases (PubMed, EMBASE, Medline), all searches should be supplemented by consulting current contents, reviews, textbooks, specialized registers, or experts in the particular field of study, and by reviewing the references in the studies found.
(E)	Journals were "hand-searched" or "manual searched" (i.e. identifying highly relevant journals and conducting a manual, page-by-page search of their entire contents looking for potentially eligible studies)

4. Was the status of publication (i.e. grey literature) used as an inclusion criterion?

(Grey literature is literature produced at all levels of government, academia, business and industry in print and electronic formats, but is not controlled by commercial publishers. Examples can be but not limited to dissertations, conference proceedings.) Here is an extra description of what grey literature is.

If it satisfies 3 of the criteria →4
If it satisfies 2 of the criteria →3
If it satisfies 1 of the criteria →2
If it satisfies 0 of the criteria →1

Criteria:

(A)	The authors should state that they searched for reports regardless of their publication type.
(B)	The authors should state whether or not they excluded any reports (from the systematic review), based on their publication status, language etc.
(C)	"Non-English papers were translated" or readers sufficiently trained in foreign language
(D)	No language restriction or recognition of non-English articles

5. Was a list of studies (included and excluded) provided?

If it satisfies 4 of the criteria →4
If it satisfies 3 of the criteria →3
If it satisfies 2 of the criteria →2
If it satisfies 1 or 0 of the criteria → 1

Criteria:

(A)	Table/list/or figure of **included** studies, a reference list does not suffice.
(B)	Table/list/figure of **excluded** studies[1] either in the article or in a supplemental source (i.e. online). (Excluded studies refers to those studies seriously considered on the basis of title and/or abstract, but rejected after reading the body of the text)
(C)	Author satisfactorily/sufficiently stated the **reason for exclusion** of the seriously considered studies.
(D)	Reader is able to **retrace** the included **and** the excluded studies anywhere in the article bibliography, reference, or supplemental source

6. Were the characteristics of the included studies provided?

If it satisfies 3 of the criteria →4
If it satisfies 2 of the criteria →3
If it satisfies 1 of the criteria →2
If it satisfies 0 criteria → 1

Criteria:

(A)	In an aggregated form such as a table, data from the original studies should be provided on the participants, interventions **AND** outcomes.
(B)	Provide the ranges of **relevant** characteristics in the studies analyzed (e.g. age, race, sex, relevant socioeconomic data, disease status, duration, severity, or other diseases should be reported.)
(C)	The information provided appears to be complete and accurate (i.e. there is a tolerable range of subjectivity here. Is the reader left wondering? If so, state the needed information and the reasoning).

7. Was the scientific quality of the included studies assessed and documented?

If it satisfies 4 of the criteria →4
If it satisfies 3 of the criteria →3
If it satisfies 2 of the criteria →2
If it satisfies 1 or 0 of the criteria → 1

Criteria:

(A)	'A priori' methods of assessment should be provided (e.g., for effectiveness studies if the author(s) chose to include only randomized, double-blind, placebo controlled studies, or allocation concealment as inclusion criteria); for other types of studies alternative items will be relevant.
(B)	The scientific quality of the included studies appears to be meaningful.
(C)	Discussion/recognition/awareness of level of evidence
(D)	Quality of evidence should be rated/ranked based on characterized instruments. (Characterized instrument is a created instrument that ranks the level of evidence, e.g. GRADE[Grading of Recommendations Assessment, Development and Evaluation.])

8. Was the scientific quality of the included studies used appropriately in formulating conclusions?

If it satisfies 4 of the criteria →4
If it satisfies 3 of the criteria →3
If it satisfies 2 of the criteria →2
If it satisfies 1 or 0 of the criteria → 1

Criteria:

(A)	The results of the methodological rigor and scientific quality should be considered in the analysis and the conclusions of the review
(B)	The results of the methodological rigor and scientific quality are **explicitly stated** in formulating recommendations.
(C)	To have conclusions integrated/drives towards a clinical consensus statement
(D)	This clinical consensus statement drives toward revision or confirmation of clinical practice guidelines

9. Were the methods used to combine the findings of studies appropriate?

If it satisfy 4 of the criteria → 4
If it satisfy 3 of the criteria → 3
If it satisfy 2 of the criteria → 2

[1] It is worth to have a brief overview of the excluded studies, since they do present relevant clinical information.

If it satisfy 1 or 0 of the following criteria → 1

Criteria:

(A)	Statement of criteria that were used to decide that the studies analyzed were similar enough to be pooled?
(B)	For the pooled results, a test should be done to ensure the studies were combinable, to assess their homogeneity (i.e. Chi-squared test for homogeneity, I^2).
(C)	Is there a recognition of heterogeneity or lack of thereof
(D)	If heterogeneity exists a "random effects model" should be used and/or the rationale (i.e. clinical appropriateness) of combining should be taken into consideration (i.e. is it sensible to combine?), or stated explicitly
(E)	If homogeneity exists, author should state a rationale or a statistical test

10. Was the likelihood of publication bias (a.k.a. "file drawer" effect) assessed?

If it satisfies 3 of the criteria →4
If it satisfies 2 of the criteria →3
If it satisfies 1 of the criteria →2
If it satisfies 0 of the criteria →1

Criteria:

(A)	Recognition of publication bias or file-drawer effect
(B)	An assessment of publication bias should include graphical aids (e.g., funnel plot, other available tests)
(C)	Statistical tests (e.g., Egger regression test).

11. Was the conflict of interest stated?

If it satisfies 3 of the criteria →4
If it satisfies 2 of the criteria →3
If it satisfies 1 of the criteria →2
If it satisfies 0 of the criteria →1

Criteria:

(A)	Statement of sources of support
(B)	No conflict of interest. This is subjective and may require some deduction or searching.
(C)	An awareness/statement of support or conflict of interest in the **primary** inclusion studies

REFERENCES

[1] Cochrane AL. Effectiveness and Efficiency: Random Reflections of Health Services. 2nd ed. London: Nuffield Provincial Hospitals Trust 1971.

[2] Littell JH, Corcoran J, Pillai V. Systematic reviews and meta-analysis. New York, NY: Oxford Univeristy Press 2008; p. 220.

[3] Chiappelli F. The science of research synthesis: a manual of evidence-based research for the health sciences. NovaScience Publisher, Inc. 2008; pp. 1-327.

[4] CRD – critical review dissemination. systematic reviews. York GB, York university press 2009.

[5] Moher D, Tetzlaff J, Tricco AC, Sampson M, Altman DG. Epidemiology and reporting characteristics of systematic reviews. PLoS Med 2007; 4(3): e78.

[6] Shojania KG, Sampson M, Ansari MT, Ji J, Doucette S, Moher D. How quickly do systematic reviews go out of date? a survival analysis. Ann Intern Med 2007; 147(4): 224-33.

[7] Savoie I, Helmer D, Green CJ, Kazanjian A. Beyond Medline: reducing bias through extended systematic review search. Int J Technol Assess Health Care 2003; 19(1): 168-78.

[8] Chiappelli F. Sustainable evidence-based decision-making. Novascience publisher, Inc. 2010.

[9] Kung J, Trinh D, Chiappelli F. Evaluating discordant systematic reviews in clinical dentistry. IADR/AADR/CADR 87th General Session and Exhibition 2009.

[10] Oxman AD. Checklists for review articles. BMJ 1994; 309: 648-51.

[11] Shea B, Dubé C, Moher D. Assessing the quality of reports of systematic reviews: the QUOROM statement compared to other tools. In: Egger M, Smith GD, Altman DG, Eds. Systematic reviews in health care: Meta-analysis in context. London: BMJ Books 2001; pp. 122-39.

[12] Sacks H, Berrier J, Reitman D, Ancona-Berk VA, Chalmers TC. Meta-analyses of randomized controlled trials. N Engl J Med 1987; 316(8): 450-55.

[13] Shea BJ, Grimshaw JM, Wells GA, *et al*. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. BMC Med Res Methodol 2007; 15: 7-10.

Received: September 24, 2009 Revised: October 03, 2009 Accepted: October 03, 2009