Case Report ■

Developing and Evaluating Criteria to Help Reviewers of Biomedical Informatics Manuscripts

ELSKE AMMENWERTH, PHD, ASTRID C. WOLFF, PHD, PETRA KNAUP, PHD, HANNO ULMER, PHD, STEFAN SKONETZKI, MSC, JAN H. VAN BEMMEL, PHD, ALEXA T. MCCRAY, PHD, REINHOLD HAUX, PHD, CASIMIR KULIKOWSKI, PHD

Abstract Peer-reviewed publication of scientific research results represents the most important means of their communication. The authors have annually reviewed a large heterogeneous set of papers to produce the International Medical Informatics Association (IMIA) *Yearbook of Medical Informatics*. To support an objective and high-quality review process, the authors attempted to provide reviewers with a set of refined quality criteria, comprised of 80 general criteria and an additional 60 criteria for specific types of manuscripts. Authors conducted a randomized controlled trial, with 18 reviewers, to evaluate application of the refined criteria on review outcomes. Whereas the trial found that reviewers applying the criteria graded papers more strictly (lower overall scores), and that junior reviewers appreciated the availability of the criteria, there was no overall change in the interrater variability in reviewing the manuscripts. The authors describe their experience as a "case report" and provide a reference to the refined quality review criteria without claiming that the criteria represent a validated instrument for quantitative quality measurement.

■ **J Am Med Inform Assoc.** 2003;10:512–514. DOI 10.1197/jamia.M1062.

Research is defined as carrying out an investigation into a subject or problem.[1] Communicating research results in recognized, peer-reviewed scientific journals is essential both to scientific progress and to individual professional advancement.[2] However, as Seglen states: "Evaluating scientific quality is a notoriously difficult problem which has no standard solution."[3] The authors' motivation to develop and evaluate quality criteria for scientific papers arose during their work in editing the *Yearbook of Medical Informatics* of the International Medical Informatics Association (IMIA).[4,5] The Yearbook appears annually and presents approximately 50 significant papers that have been published during the previous year. Its aim is to give a broad overview of the latest significant research activities in the field of health and medical informatics.

During the selection process, approximately 10,000 medical informatics papers published annually and listed in MEDLINE are reviewed and filtered to retain about 50 papers for publication in the Yearbook. Eight managing editors, each assigned to different subfields, first preselect papers for review; those (approximately) 150 papers are reviewed by two external international experts, by the two editors of the Yearbook, and by the responsible managing editor. A purely quantitative review scale is used for the final review. An analysis of the external reviews of 118 papers preselected for the IMIA Yearbook 2001[6] indicated a wide range of variability in "expert" scoring, with approximately one-third of the papers showing a difference in quantitative scores of 20% or more among the two external reviewers.

The editors and managing editors attempted to refine the main review criteria with the goal of decreasing rater variability and improving quality of reviews. The authors of the current report conducted a review of the relevant literature; developed new, refined review criteria; and conducted a small randomized controlled trial using the new criteria to assess variability and reviewers' satisfaction with the new criteria. The trial indicated that the new criteria were not useful, in an absolute sense, to create better agreement among reviewers. This is not surprising in that reviewers with different expertise and different experiences will view "relevant" aspects of a paper from unique perspectives.

Affiliations of the authors: Research Group Assessment of Health Information Systems, University for Health Informatics and Technology Tyrol (UMIT), Innsbruck, Austria (EA); Department of Medical Informatics, University of Heidelberg, Germany (ACW, PK, SS); Institute for Biostatistics and Documentation, University of Innsbruck, Austria (HU); Erasmus University, Rotterdam, The Netherlands (JHV); National Library of Medicine, Bethesda, Maryland (ATM); Institute for Health Information Systems, University for Health Informatics and Technology Tyrol, Innsbruck (UMIT), Austria (RH); Department of Computer Science, Rutgers, The State University of New Jersey, New Brunswick, New Jersey (CK).

The authors are Editors and Managing Editors of the IMIA Yearbook 2001. The authors thank Martina Hutter, Steven Huesing, Thomas Kleinöder, and the Schattauer Publishing Company for their support in publishing the IMIA Yearbook. They also want to thank the fellow Managing Editors A. Bohne, K. Ganser, C. Maier, A. Michel, V. Mludek, and R. Singer for their discussions and comments on this paper as well as the 16 reviewers (S. Abel, B. Baumgarten, A. Bess, B. Brigl, T. Bürkle, E. Finkeissen, S. Garde, E. Lang, F. Leiner, F. Phillip, J. Pilz, U. Prokosch, M. Schwabedissen, R. Weber, T. Wendt, T. Wetter) for their participation in the study. Thanks also to Frieda Kaiser for her support and the anonymous reviewers for their fruitful comments on an earlier version of this paper.

Correspondence and reprints: Asst.-Prof. Dr. Elske Ammenwerth, Research Group Assessment of Health Information Systems, University for Health Informatics and Technology Tyrol (UMIT), Innrain 98, 6020 Innsbruck, Austria; e-mail: <elske.ammenwerth@umit.at>.

Received for publication: 11/21/01; accepted for publication: 05/06/03.

The goal of this article is to report the authors' attempt to define quality manuscript review criteria and to share what was learned about their application during reviews.

Development of Refined Quality Criteria for Paper Reviews

The authors first reviewed the available literature on review criteria for scientific articles. For example, Gunn[7] discussed problems with the quality of electronically published clinical guidelines, such as low quality and irrelevance. Elliott et al.[9] presented guidelines for reviewing qualitative research, and Jefferson et al.[10] assessed whether *BMJ* guidelines for reviewing economics submissions influenced the quality of submitted and published manuscripts in this area. Also, a standard was developed for measuring quality of publications regarding randomized controlled clinical trials (the Consolidated Standards of Reporting Trials, or CONSORT system).[11] CONSORT implemented a checklist containing 21 items, covering manuscripts' methods, results, and discussion sections.[11] The German Research Association, among others, has published general guidelines for good scientific practice[12]; their recommendation 12 covers (co-)authorship, completeness of presentation of research results, and correct citation of previous work of other researchers.

Many scientific medical informatics journals provide information regarding their own quality criteria via instructions for authors and guidelines for reviewers (e.g., *BMJ*[13] and *JAMIA*[14]). Review criteria of medical journals (such as *BMJ*) are not always directly relevant to biomedical informatics publications.

The authors' literature review showed that no previously published quality checklist met the objective of providing a comprehensive list of refined quality criteria useful for reviewing all scientific papers in medical informatics.

The authors chose a top-down approach to refining the previously used IMIA Yearbook main review criteria (with sections for significance, quality of scientific content, originality and innovativeness, coverage of related literature, and organization and clarity of presentation). Analysis of available literature[11,13–16] and the authors' own experiences as reviewers provided the basis for the refined criteria, which then were discussed and revised by IMIA Yearbook editors and managing editors in an iterative process.[17]

The revised review criteria developed by this methodology are presented on the Web pages of the IMIA Yearbook at <http://www.yearbook.uni-hd.de>. The revised criteria included five quality categories with 15 subgroups, totaling approximately 80 general questions, with approximately 60 additional questions for specific subtypes of articles. Table 1 (available as an online data supplement at www.jamia.org) highlights some of the differences between the review criteria of *BMJ*, *JAMIA*, and CONSORT and compares them with the authors' refined quality criteria.

Evaluation of Revised Quality Criteria and Lessons Learned

After developing and elaborating the revised quality criteria, the authors conducted a randomized trial, comparing the new review criteria with "standard" previous review methods (see Appendix A in an online data supplement at www.jamia.org for methods and results of that study). The randomized study failed to show an effect of the revised criteria on interrater concordance. Nevertheless, the authors found the criteria to be helpful in the following manner.

1. The small-scale evaluation of the refined quality criteria showed that the reviewers' absolute quality ratings fell significantly (lower scores of merit) on papers they reviewed while applying the new criteria. Reviewers commented that the refined quality criteria helped to increase their awareness of all quality criteria, so that they more easily identified weaknesses in the reviewed papers, justifying lower ratings. Stricter ratings may be a desired effect, reflecting improved review quality—especially for less experienced reviewers. More experienced reviewers may already have most of the criteria in mind.
2. An observed increased time required to apply the refined quality criteria was not surprising. In the study, all reviewers had to grade each of the 140 individual review criteria for every paper reviewed, so that the study could be certain that the refined criteria had been applied. During "normal" reviews, it would be expected that refined criteria would serve as useful information for novice reviewers and as infrequently used but available reference material for more experienced reviewers. The authors believe that the increased time required to use the refined criteria may not reduce their usefulness as long as the criteria are available as a reference and not as an absolute requirement during review (a subject for future study).
3. Different reviewers of equal stature and ability will always judge the manuscript differently due to their varied scientific backgrounds, their variable familiarity with similar projects, and their variegated knowledge of the authors and the authors' prior work. It is therefore not surprising that the authors' randomized study found no reduction in variation between the reviewers after using the refined quality criteria.
4. The refined quality criteria do not constitute an instrument that can be used to obtain an objective quantitative assessment from reviewers and cannot lead to a "definitive" quality score. Rather, the criteria provide a list of important considerations that can help to support thorough reviews of scientific papers (even though the results will be different due to different backgrounds of the reviewers).
5. The reviewer (or author) of a paper can use the refined criteria as a reference in case of questions, for aspects to consider when judging originality or significance of a paper. Refined quality criteria may help novice reviewers to think more thoroughly about reviews (thus perhaps leading to better reviewers). The criteria may help in assessing how and why to arrive at certain decisions.
6. The list of refined quality criteria is not designed to be exhaustive and will be continually revised annually by the IMIA Yearbook editorial staff.

The quality of reporting in a paper does not automatically reflect the quality of the underlying research project it describes. But a good paper makes it easier to assess the quality of a research project. Publication quality is an important aspect of research quality.[15]

The authors hope that the refined review criteria will be helpful for authors of scientific papers in medical informatics

and for reviewers and editors to come to a balanced and more explicit assessment of the quality of medical informatics papers. The journal *Methods of Information in Medicine* has already adopted a draft of the refined quality criteria for its review guidelines.[18]

References ■

1. Pons Collins Dictionary of the English Language. Glasgow: HarperCollins, 1998.
2. de Vries J. Peer review: the holy cow of science. In: Fredriksson E (ed). A Century of Science Publishing Amsterdam: IOS Press, 2001, pp 231–44.
3. Seglen PO. Why the impact factor of journals should not be used for evaluating research. BMJ. 1997;314:498–502.
4. van Bemmel J, McCray A. Yearbook of Medical Informatics (annual edition). Stuttgart: Schattauer, 1992–2000.
5. Haux R, Kulikowski C. Yearbook of Medical Informatics (annual edition). Stuttgart: Schattauer, since 2001.
6. Ammenwerth E, Knaup P, Maier C, et al. Digital libraries and recent medical informatics research—Findings from the IMIA yearbook of medical informatics 2001. Methods Inf Med. 2001;40:163–7.
7. Gunn IP. Evidence-based practice, research, peer review, and publication. CRNA (clinical forum for nurse anesthetists). 1998;9:177–82.
8. Boyer C, Selby M, Scherrer J-R, Appel R. The health on the net code of conduct for medical and health websites. Comput Biol Med. 1998;28:603–10.
9. Elliott R, Fischer CT, Rennie DL. Evolving guidelines for publication of qualitative research studies in psychology and related fields. Br J Clin Psychol. 1999;38(pt 3): 215–29.
10. Jefferson T, Smith R, Yee Y, et al. Evaluating the BMJ guidelines for economic submissions: prospective audit of economic submissions to *BMJ* and *The Lancet*. JAMA. 1998; 280:275–7.
11. International Committee of Medical Journal Editors. Uniform requirements for manuscripts submitted to biomedical journals. JAMA. 1997;277:927–34.
12. DFG. German Research Association: Proposals for safeguarding good scientific practice. Weinheim: Wiley-VCH, 1998.
13. BMJ. Checklist for editors and peer reviewers. <http://bmj.com/advice/>. Accessed October 2002.
14. Journal of the American Medical Informatics Association. Instructions for authors. <http://www.jamia.org/misc/ifora.shtml> Accessed October 2002.
15. Köbberling J. The quality of German medical journals [in German]. Dtsch Med Wschr. 2000;125:1106–8.
16. Paice E. How to write a peer review. Hosp Med. 2001;62: 172–5.
17. Kulikowski C, Ammenwerth E, Bohne A, et al. Medical imaging informatics and medical informatics: opportunities and constraints—findings from the IMIA Yearbook of Medical Informatics 2002. Methods Inf Med. 2002;41:183–9.
18. Methods of Information in Medicine. Instructions for authors. Accessed October 2002. <http://www.schattauer.de/zs/startz.asp?load=/zs/methods/richtl.asp>.

JAMIA M1062 ONLINE DATA SUPPLEMENT APPENDIX ONE:

Effects of the Refined Quality Criteria on Reviews: A Study

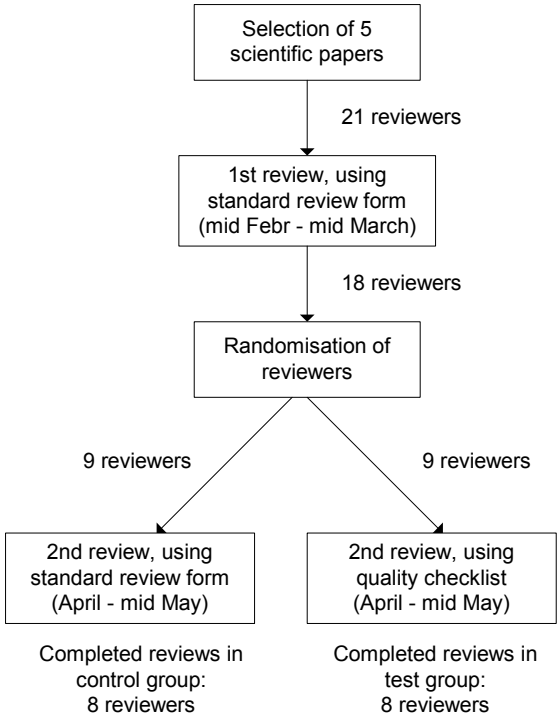
1. Study design

In order to evaluate the effects of the refined quality criteria on reviews, we conducted a randomized controlled trial. The aim was to answer the following three questions:
Q1 Does the variation between the reviewers change when the refined quality criteria are used?
Q2 Do the quantitative judgments of the reviewers change when the refined quality criteria are used?
Q3 Do the reviewers find the refined quality criteria useful to support reviews?
The study took place between February and May 2002. Twenty-one medical informatics researchers working in the area of information systems agreed to participate as reviewers in this study. Five papers were selected to be included in the study, taken as a sample from the about 60 candidate papers for the three information systems sections of the IMIA Yearbook 2002 [17].
The study was designed as a randomized controlled trial. The reviewers in the test group were asked to review each of the 5 papers twice: First with the usual 1-page evaluation form of the IMIA Yearbook with the five main quality criteria, and then again with the refined quality criteria. In order to be able to attribute any effect to the refined quality criteria, a control group of reviewers was defined which also evaluated each paper twice, but taking the 1-page evaluation form each time.
The distribution of reviewers to either the test group or control group was done randomly, stratified for their review experience. The washout time between the first and the second review was set to about 8 weeks. We considered this period sufficient to minimize the memory of details of the first review. In order to support this, all reviewed papers were re-collected after the 1st review, and the reviewers were asked not to keep a copy of the 1st review ratings. In order to guarantee that the reviewers of the test group really used the refined quality criteria during the second review, they were asked to check each item individually and note agreement or disagreement, before giving their overall rating for each category.
To answer Q1 on a change in the variation by the refined quality criteria, the t-test for paired samples was used to compare mean coefficients of variations between the first and second review in the test group.
To answer Q2 on a change in the mean ratings by the refined quality criteria, a four-way analysis of variance (ANOVA) with repeated measurements was used.
In both situations, to assure the applicability of the t-test and the analysis of variance, the Kolmogorov-Smirnov-Test with Lilliefors was used to check for normal distributions of the given ratings.
To answer Q3, the reviewers were asked to judge the usefulness of the refined quality criteria (only in the test group, after the 2nd review, using a Likert scale and open-ended questions).
All reviewers were also asked to document the time needed to complete the reviews, and to indicate the review experience (after the 1st review, using a Likert scale and asking for the years of review experience).

2. Execution of the study

The five selected papers covered various topics such as data mining, pharmacy system, clinical guidelines, computer-based reminders, and medication errors. Figure 1 shows the execution of the study. From the 21 reviewers who agreed to participate, 18 completed the first review. They were then randomized into the test group (9 researchers) and into the control group (9 researchers). 8 reviewers in each group completed the 2nd review. The reviewers who left the study gave as reason insufficient time to complete reviews.
From the eight reviewers in each group, five stated they had relatively little experience in reviewing papers (0 - 2 years of experience), and 3 stated that they were more experienced reviewers (5 – 15 years of experience). None of them had participated in the development of the refined quality criteria. The mean number of days between the first and second review of the same papers was 64 days ± 12 days, or about 2 months.

Figure 1: Flowchart of the study implementation for evaluating the effects of the quality checklist. Overall, 16 reviewers completed the study (8 in the test group, 8 in the control group).



3. Study results

i. Change of variation between reviewers

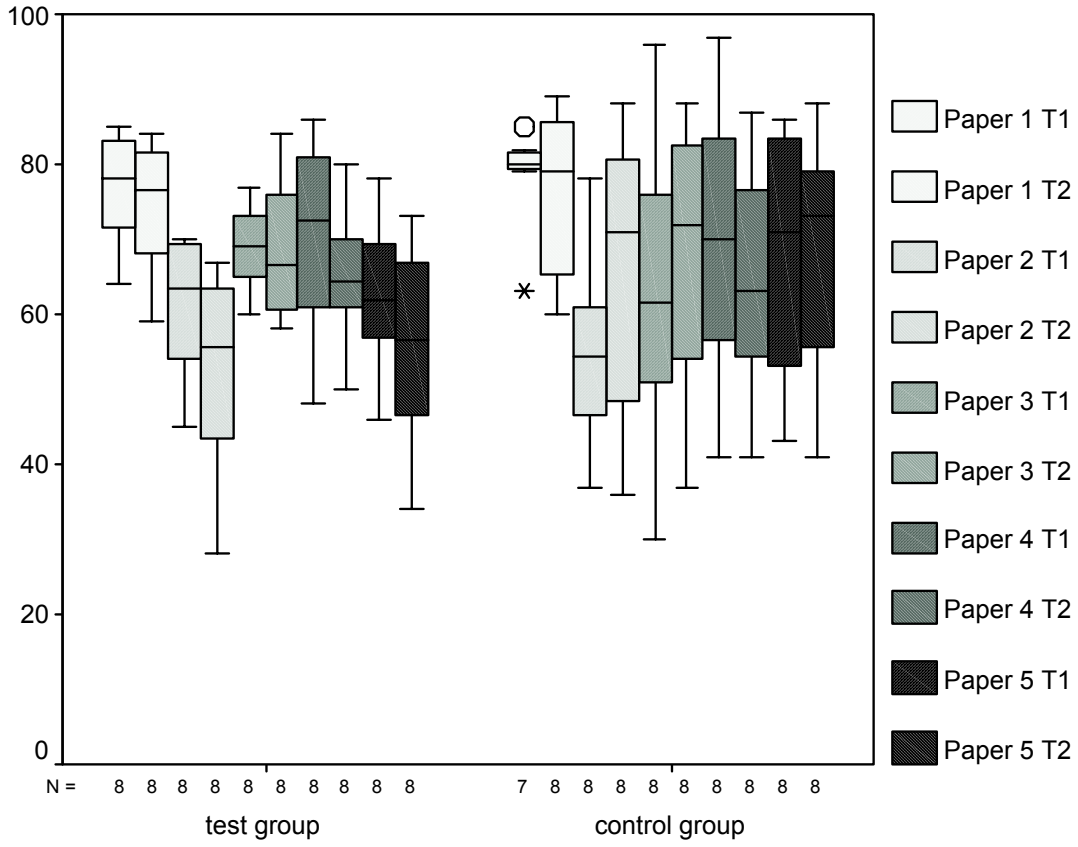
The variation index was chosen as an index for the variability of reviews. It was calculated for each question as standard deviation divided through the mean. In the control group, the variation index was 0.29 ± 0.12 at the 1st review, and 0.27 ± 0.09 in the 2nd review. In the test group, the variation index was 0.23 ± 0.08 and 0.24 ± 0.10, respectively.
Since it could be shown with the Kolmogorov-Smirnov-Test with Lilliefors correction that the variation indices were approximately normally distributed, the paired t-test could be used to check the hypothesis of a change in variation. The hypothesis could not be rejected. Thus, no reduction of variation between the reviewers in the test group could be confirmed by using the refined quality criteria.

ii. Change of mean ratings of papers

Figure 2 shows the mean ratings and distribution for each paper in the test group and in the control group. All papers have already been published and have thus been peer-reviewed, therefore it is not surprising that nearly all ratings are higher than 60 (from 100).

The Kolmogorov-Smirnow-Test with Lilliefors correction showed that the ratings are approximately normally distributed. Therefore, a 4-factor analysis of variance with repeated measurements could be conducted, using paper, time and question as within-subject factors, and group as between-subject factor. The results showed a significant interaction between paper, review time and group ($p = 0.027$), and between review time and group ($p = 0.034$). Post hoc analyses showed that the mean overall rating in the test group was significantly reduced from 68.0 ± 6.8 in the 1st review to 63.3 ± 4.1 in the 2nd review ($p = 0.001$), with the mean reduction being about 5 points (equivalent to 5%). In the control group, no significant changes could be found (66.2 ± 12.6 vs. 68.4 ± 14.0; $p = 0.485$).

Figure 2: Distribution of ratings for each of the 5 papers in the test group (8 reviewers) and in the control group (8 reviewers). T1 = 1st review, T2 = 2nd review, using box-whisker-plots. The maximum possible rating score is 100, indicating the highest quality, the lowest quality rating being 0.



Thus, the hypotheses of a change in mean ratings could not be rejected for the main factor effects group, review time and paper, but it could be rejected for the interaction of review time and group as well as for the interaction of review time, group and paper. The significant interaction between review time and group was constituted by lower (stricter) ratings in the test group and no change of the ratings in the control group.

iii. Usefulness of refined quality criteria in the opinion of reviewers

Table 2 shows the time needed to complete the review, as documented by the reviewers. In the control group the time needed to review the papers decreased by about 1/3 during the 2nd review, while in the test group, the review time increased by about 1/3. Both results are not surprising: The time needed to read and rate a paper is certainly lower when the paper has already been read and reviewed earlier, and higher when using the extended refined quality criteria instead of the 1-page standard criteria. Altogether, the mean review time was increased by about 50% when the refined quality criteria is thoroughly used.

Table 2: Mean time and standard deviation (in minutes) to terminate the review of the five papers during the 1st review and the 2nd review. The test group used the refined quality criteria during the 2nd review, otherwise the standard 1-page evaluation form was used.

	1st review	2nd review
Control group	32.6 ± 12.2	22.8 ± 10.2
Test group	25.4 ± 10.9	33.8 ± 17.7

Table 3 showed how the reviewers of the test group judged the usefulness of the refined quality criteria. In the free comments, the three more experienced reviewers remarked that the use of the refined quality criteria takes too much time (n=3), that some criteria cannot be applied to all papers (n=1), and that certain types of papers (such as innovative papers) cannot adequately be judged with this list (n=2). On the positive side, one experienced reviewer commented that subjective opinions can now be better justified.

Table 3: Usefulness of the refined quality criteria, as seen by the 8 reviewers of the test group, after the 2nd review, on a 5-point Likert scale (-- = absolutely not, - = rather not, -/+ = maybe, + = rather yes, ++ = absolutely yes).

		--	-	-/+	+	++
Less experienced reviewers (n=5)	Felt that list supported me in the review				2	3
	Will use list to support further reviews			1	4	
More experienced reviewers (n=3)	Felt that list supported me in the review		1	1	1	
	Will use list to support further reviews		2		1	

The five less experienced reviewers stated that the refined quality criteria support reviews when review experience is low (n=2), that it helps to become conscious of the criteria (n=2), to justify more negative ratings (n=1), and that it supports a structured review process (n=1).

JAMIA M1062 ONLINE DATA SUPPLEMENT Table 1:

Structure and size of the review criteria of BMJ, JAMIA, CONSORT and of the proposed refined quality criteria.

	BMJ	JAMIA	CONSORT	Proposed refined quality criteria
Significance, scientific impact	16 criteria (as part of rejection checklist)	(part of criteria for specific types of papers)		5 criteria
General quality of content	12 criteria (as part of rejection checklist) 6 criteria (as part of scientific reliability)	(part of criteria for specific types of papers)		25 criteria
Additional criteria for certain types of papers	12 criteria for qualitative research 11 for general statistics 26 for RCT 35 for health economics papers 23 for clinical management guidelines	5 criteria for review papers 3 criteria for viewpoint papers 4 criteria for application reports 6 criteria for model formulation papers 6 criteria for research papers	21 criteria for RCT reports	10 criteria for empirical investigations (incl. RCT) 7 for qualitative research papers 5 for methodological papers 7 for application reports 7 for systematic reviews 6 for viewpoint papers
Originality	1 criteria	(part of criteria for specific types of papers)		3 criteria
Coverage of literature	1 criteria (as part of scientific reliability)	(part of criteria for specific types of papers)		4 criteria
Organization of paper		(part of criteria for specific types of papers)		14 criteria