ANZ J. Surg. 2003; **73**: 712–716

ORIGINAL ARTICLE

METHODOLOGICAL INDEX FOR NON-RANDOMIZED STUDIES (*MINORS*): DEVELOPMENT AND VALIDATION OF A NEW INSTRUMENT

KAREM SLIM,* EMILE NINI,* DAMIEN FORESTIER,* FABRICE KWIATKOWSKI,† YVES PANIS‡ AND JACQUES CHIPPONI*

**Department of General and Digestive Surgery, Hôtel-Dieu, Clermont-Ferrand, †Department of Statistics, Centre Jean-Perrin Clermont-Ferrand and ‡Department of Digestive Surgery, Hôpital Lariboisière, Paris, France*

Background: Because of specific methodological difficulties in conducting randomized trials, surgical research remains dependent predominantly on observational or non-randomized studies. Few validated instruments are available to determine the methodological quality of such studies either from the reader's perspective or for the purpose of meta-analysis. The aim of the present study was to develop and validate such an instrument.

Methods: After an initial conceptualization phase of a methodological index for non-randomized studies (MINORS), a list of 12 potential items was sent to 100 experts from different surgical specialities for evaluation and was also assessed by 10 clinical methodologists. Subsequent testing involved the assessment of inter-reviewer agreement, test-retest reliability at 2 months, internal consistency reliability and external validity.

Results: The final version of MINORS contained 12 items, the first eight being specifically for non-comparative studies. Reliability was established on the basis of good inter-reviewer agreement, high test-retest reliability by the κ-coefficient and good internal consistency by a high Cronbach's α-coefficient. External validity was established in terms of the ability of MINORS to identify excellent trials.

Conclusions: MINORS is a valid instrument designed to assess the methodological quality of non-randomized surgical studies, whether comparative or non-comparative. The next step will be to determine its external validity when used in a large number of studies and to compare it with other existing instruments.

Key words: comparative study, methodology index, non-randomized study.

Abbreviation: MINORS, methodological index for non-randomized studies.

INTRODUCTION

Although surgeons are now conducting an increasing number of randomized trials,[1] most of the available evidence in surgery comes from non-randomized studies, both comparative and non-comparative. Indeed surgical research remains an example of a situation where randomization is not always possible or feasible.[2] Beyond large randomized trials, systematic reviews are an important way to answer questions in surgery. However, the systematic review or meta-analysis of studies other than randomized trials may be difficult because combining the results of observational studies of heterogeneous quality could be highly biased.

Observational studies include comparative studies such as case-control and cohort designs, and patient series which may or may not involve comparisons between two or more groups.

Several papers have discussed the methodology of meta-analyses of observational studies[3,4] and checklists have been proposed but not formally validated.[5] Downs and Black used clinimetric criteria to develop a checklist which was applicable to both randomized and non-randomized studies without distinction.[6] The aim of the present study was to develop and validate a methodological index for non-randomized studies (MINORS) which could be used by readers, manuscript reviewers or journal editors to assess the quality of such studies.

K. Slim MD, FACS; **E. Nini** MD; **D. Forestier** MD; **F. Kwiatkowski** PhD; **Y. Panis** MD, PhD; **J. Chipponi** MD, PhD.

Correspondence: K. Slim, Department of General and Digestive Surgery, Hôtel-Dieu Boulevard Leon Malfreyt, F-63058 Clermont-Ferrand, France. Email: kslim@chu-clermontferrand.fr

Accepted for publication 13 May 2003.

METHODS

Conceptualization phase

After reviewing the literature on quality assessment of randomized trials and discussing the particular features of non-randomized studies, a panel of eight practising surgeons selected 12 items to be considered for inclusion in MINORS. These items were chosen because of their ability to characterize the methodological and scientific value of published articles. Seven items were selected for assessment of non-comparative studies and five for use with comparative studies. The list of 12 items was then sent to 100 surgeons throughout France who had clinical research expertise in different specialities, including digestive, cardiovascular and thoracic surgery, gynaecology, otorhinolaryngology, orthopaedics, urology, neurosurgery, and ophthalmology. They were asked to score the ability of each item to assess the quality of a given study using a 7-point-scale, according to the method proposed by Oxman and Guyatt.[7] The mean score for each item was then compared with that of every other item to see whether there were any significant differences. Subsequently each item was scored from 0 to 2; 0 indicating that it was not reported in the article evaluated, 1 indicating that it was reported but inadequately, and 2 indicating that

it was reported adequately. The form also included a section allowing the surgeons to suggest additional items.

Assessment of face validity and content validity

To determine whether MINORS items appeared appropriate and whether they covered all important considerations relevant to the methodology of non-randomized studies, a revised list of items was sent to 10 French clinical methodologists for assessment on 13 credibility criteria according to the method proposed by Feinstein.[8]

Clinimetric testing of MINORS

Inter-reviewer agreement
To test the consistency of MINORS between reviewers, a random sample of published non-randomized studies, both comparative and non-comparative, was selected from among several specialties. For this purpose a Medline search was undertaken using the MeSH 'surgery' and limits by publication type (clinical trial *not* randomized controlled trial) for the year 2001. A numerical list of original articles was then established and 80 articles were selected randomly. The title, authors' names, institutional affiliation and journal identity were removed. These articles were then assessed by two independent reviewers with different methodological expertise (one junior and one senior surgeon) using the revised version of MINORS.

Test-retest reliability
Two months after the first assessment, a randomly selected sample of 30 articles was scored again by the junior surgeon without reference to his first assessment.

Internal consistency
This evaluation indicated whether the items were related to one another and worked together in a similar manner in assessing the quality of articles.

Validity
The power of MINORS to differentiate between excellent, fair or poor studies was examined by selecting a random sample of 15 excellent randomized controlled trials. These articles were chosen as the gold standard against which to assess the external validity of MINORS on the basis that they had all been published in three major journals which had adopted the CONSORT Statement,[9] (namely *British Medical Journal*, *British Journal of Surgery and The Lancet*). These articles were then scored according to MINORS and the results compared with a selected group of the 15 best-scored comparative studies from the sample of 80 described previously. The reviewer was blinded as to the source of the 15 randomized trials.

Statistical analyses

Agreement between reviewers was measured by the κ-coefficient (unweighted model) with a value greater than 0.4 being accepted as satisfactory.[10] Global scores were obtained by summing all the item scores**.** Results were expressed as means (standard deviations). The matched pairs *t*-test was used to compare mean global scores between reviewers. Internal consistency was assessed by the calculation of Cronbach's α-coefficient.[11] A value of $P < 0.05$ was considered statistically significant.

RESULTS

Content and face validity

Expert phase
Ninety of the 100 experts returned a completed form. Table 1 summarizes the scores of the 12 items included in the first version of MINORS. No item was scored less than five. Furthermore there was no difference between the different specialities. The experts suggested no additional methodological items apart from a modification of item 11. As a result a supplementary sentence was incorporated in that item in the revised version of MINORS relating to the size of non-comparative studies. Item 11 thus became relevant to both comparative and non-comparative studies. The revised version of MINORS included 12 items: the first subscale of eight items related to non-comparative studies whereas all 12 items were relevant to comparative studies (Table 2).

Revision phase
Because there was no statistical difference between the mean item scores as evaluated by the experts, the items were not weighted and the scoring was simplified to a 3-point scale from 0 to 2. If one considers that MINORS involves eight items for non-comparative studies and 12 items for comparative studies and that the maximum item score is 2, the ideal global score would be 16 for the non-comparative studies and 24 for the comparative studies.

Methodologist phase
All 10 methodologists completed their assessment and scored the final version favourably, all item mean scores being above 4.5 on a 7-point scale (Table 3).

Table 1. Assessment of items in the first version of MINORS by 90 experts in several surgical specialities using a scale from 0 to 7

Item	Median	Mean (SD)
1. A stated aim of the study	7	6.6 (0.7)
2. Inclusion of consecutive patients	6	5.8 (1.1)
3. Prospective collection of data	6	5.5 (1.2)
4. Endpoint appropriate to the study aim	6	6.3 (0.8)
5. Unbiased evaluation of endpoints	5	5.4 (1.2)
6. Follow-up period appropriate to the major endpoint	6	6.2 (0.8)
7. Loss to follow up not exceeding 5%	6	5.5 (1.2)
And in the case of comparative studies		
8. A control group having the gold standard intervention	6	6.0 (1.1)
9. Contemporary groups	6	5.5 (1.4)
10. Baseline equivalence of groups	6	6.1 (0.9)
11. Prospective calculation of the sample size	6	5.5 (1.3)
12. Statistical analyses adapted to the study design	6	6.3 (0.8)

SD, standard deviation.

Inter-reviewer agreement, test-retest reliability and internal consistency reliability

There were 26 comparative and 54 non-comparative studies in the random sample assessed by the two reviewers. Table 4 summarizes the correlation between the scores of the reviewers. Agreement between the reviewers was considered satisfactory for all items. The mean global scores on a scale from 0 to 24 were, respectively, 13.93 (0.35) for the junior surgeon and 12.98 (0.54) for the senior surgeon. This difference was statistically significant ($P < 10^{-7}$) but corresponds to only 0.95 of a global score point. The mean global scores, which ranged between 0 and 20 in this agreement assessment, did not differ significantly between the comparative and non-comparative studies ($P = 0.11$).

The assessment of test-retest reliability showed a satisfactory correlation between the original and repeated scoring after a 2-month interval. The mean global score decreased significantly from 13.91 (3.3) at the first test to 12.28 (3.6) at the second ($P < 0.0001$). The internal consistency reliability of MINORS was high with a global α-value of 0.73. This demonstrated that all items worked in a complementary and coherent manner.

Validity

The 15 gold-standard randomized trials had a mean global score of 23.1. The comparison between the score of these randomized trials and that of the 15 best comparative non-randomized studies (19.8) showed a significant difference ($P = 0.00001$) in favour of the randomized trials.

DISCUSSION

This index for the assessment of non-randomized studies was developed by a group of surgeons because of the problems faced by clinicians as to the lack of randomized surgical trials and the large number of observational studies in surgery. To apply the principles of evidence-based medicine to clinical practice requires a method for assessing the quality of published data.

Table 3. Credibility criteria assessed by 10 clinical methodologists on a 7-point scale

Criterion	Mean	SD (range)
1. Wide applicability	5.5	0.5 (5–6)
2. Use by various groups	4.8	0.8 (4–6)
3. Clarity and simplicity	5.1	0.8 (4–6)
4. Adequate instructions	5.0	1.0 (4–7)
5. Information available	4.9	1.2 (3–7)
6. Need for subjective decision	4.7	0.9 (4–7)
7. Likelihood of bias	4.8	1.2 (3–7)
8. Single domain	5.1	0.9 (4–7)
9. Redundant items	5.6	0.8 (4–7)
10. Comprehensiveness	5.1	0.7 (4–6)
11. Item weights	5.4	1.0 (3–7)
12. Number of response options	5.4	1.1 (4–7)
13. Discrimination power	5.1	0.5 (4–6)

SD, standard deviation.

Table 2. The revised and validated version of MINORS

Methodological items for non-randomized studies	Score†
1. **A clearly stated aim:** the question addressed should be precise and relevant in the light of available literature	
2. **Inclusion of consecutive patients**: all patients potentially fit for inclusion (satisfying the criteria for inclusion) have been included in the study during the study period (no exclusion or details about the reasons for exclusion)	
3. **Prospective collection of data**: data were collected according to a protocol established before the beginning of the study	
4. **Endpoints appropriate to the aim of the study**: unambiguous explanation of the criteria used to evaluate the main outcome which should be in accordance with the question addressed by the study. Also, the endpoints should be assessed on an intention-to-treat basis.	
5. **Unbiased assessment of the study endpoint**: blind evaluation of objective endpoints and double-blind evaluation of subjective endpoints. Otherwise the reasons for not blinding should be stated	
6. **Follow-up period appropriate to the aim of the study**: the follow-up should be sufficiently long to allow the assessment of the main endpoint and possible adverse events	
7. **Loss to follow up less than 5%**: all patients should be included in the follow up. Otherwise, the proportion lost to follow up should not exceed the proportion experiencing the major endpoint	
8. **Prospective calculation of the study size**: information of the size of detectable difference of interest with a calculation of 95% confidence interval, according to the expected incidence of the outcome event, and information about the level for statistical significance and estimates of power when comparing the outcomes	
Additional criteria in the case of comparative study	
9. **An adequate control group**: having a gold standard diagnostic test or therapeutic intervention recognized as the optimal intervention according to the available published data	
10. **Contemporary groups**: control and studied group should be managed during the same time period (no historical comparison)	
11. **Baseline equivalence of groups**: the groups should be similar regarding the criteria other than the studied endpoints. Absence of confounding factors that could bias the interpretation of the results	
12. **Adequate statistical analyses**: whether the statistics were in accordance with the type of study with calculation of confidence intervals or relative risk	

†The items are scored 0 (not reported), 1 (reported but inadequate) or 2 (reported and adequate). The global ideal score being 16 for non-comparative studies and 24 for comparative studies.

Table 4. List of 12 items of the definitive MINORS. Inter-reviewer correlation on a random sample of 80 articles and test-retest reliability on a random sample of 30 articles.

Methodological item for non-randomized studies	κ-coefficient for inter-reviewer agreement (SD)†	κ-coefficient for test-re-test reliability (SD)
1. A clearly stated aim	0.87 (0.07)	0.89 (0.11
2. Inclusion of consecutive patients	0.78 (0.06)	0.83 (0.09)
3. Prospective collection of data	0.79 (0.06)	0.82 (0.09)
4. Endpoints appropriate to the aim of the study	0.56 (0.09)	0.76 (0.12)
5. Unbiased assessment of the study endpoint	0.61 (0.08)	0.61 (0.13)
6. Follow-up period appropriate to the aim of the study	0.61 (0.08)	0.59 (0.11)
7. Loss to follow up less than 5%	0.69 (0.08)	0.74 (0.12)
8. Prospective calculation of the study size	1.00	1.00
Additional criteria in the case of comparative studies		
9. An adequate control group	0.86 (0.09)	1.00
10. Contemporary groups	0.79 (0.14)	0.61 (0.31)
11. Baseline equivalence of groups	0.87 (0.09)	1.00
12. Adequate statistical analyses	0.66 (0.14)	0.75 (0.22)

†A κ-coefficient of >0.4 was considered satisfactory.

This is an important consideration for the 'consumers' of clinical research. Our initial aim was to develop and validate an index which would be simple to use both by readers of published articles and reviewers of manuscripts submitted for publication, and be of sufficient sensitivity for use in meta-analysis of non-randomized studies. To achieve this we followed the recognized principles of scale construction[12] using a rigorous methodology. The results of the present study show clearly that the instrument we have developed has good reliability, internal consistency and validity. The high response rate from experts and the limited number of items used, suggest that MINORS is easy to apply. Its simplicity and objectivity is also demonstrated by its acceptability to surgeons having sound methodological expertise. Although the difference between the scores of the senior and the junior reviewers was statistically significant, its actual relevance was low as the difference did not exceed 1 point.

Similarly, the assessment of test-retest reliability showed a good correlation over an interval of 2 months. The reviewer scored perhaps more severely on the second occasion, which suggests greater expertise with further experience, but the difference was too small (1.6) to be important. Nevertheless this feature may need to be investigated further.

Instead of weighting, we chose to score the items from 0 to 2 according to whether they were reported or not and adequate or not. Weighting of items requires further investigation as we have no gold standard method to evaluate the relative importance of a given methodological item. In the light of the available literature, the most appropriate method of weighting would be based on consensus development among experienced epidemiologists before designing a large study to validate their conclusions. Few attempts have been made to estimate the respective values of some methodological items.[13] Furthermore the most significant findings regarding the weighting of items have been specifically related to randomized double blind studies. One could assume that the rationale for weighting in randomized trials can be extrapolated to non-randomized studies. However this needs to be confirmed by further investigations, especially in the field of surgery. Furthermore, the item weights could differ according to the type of study. For example unbiased evaluation of endpoints is important for functional disorders whereas the length of follow up and loss to follow up are important for hernia or cancer surgery. Currently, however, there is no sound evidence for the differential weighting of items in methodological indices or checklists for non-randomized studies.

Downs and Black[6] reported a checklist applicable to both randomized and non-randomized trials. It involved 27 items concerning external validity, bias, confounding factors, statistical power and reporting; however, the number of items and differences in scoring systems between items increased complexity and user burden. Several items were related to reporting and thus were not directly concerned with the methodological quality of a study. Also in their study, the period between the test and re-test was only 2 weeks and the reviewers were similar to one-another in their level of methodological skill. Furthermore their instrument was a checklist and was not developed as an index for scoring studies.

An important aspect of MINORS is its external validity; that is, its ability to identify high quality studies, which was established by comparison with the current standard for randomized trials, namely the CONSORT Statement. Since MINORS does not differentiate between randomized and non-randomized studies and includes several items derived from indices focusing on the quality of randomized trials, the fact that a given study has a randomized design is not sufficient to achieve a high score. MINORS was not developed specifically to assess the quality of randomized trials; however, we considered the randomized trial to be the best example of comparative studies and assumed that MINORS should be able to distinguish between different comparative studies. MINORS satisfied that expectation and clearly confirmed that a good randomized trial scores higher than a good non-randomized comparative study. The ability of MINORS to recognize the poor or fair quality of non-comparative studies is suggested in our study, but this needs to be further evaluated by comparison with the Downs and Black checklist.[6] This comparison will be the subject of a future study to develop a reliable standardized instrument for assessing the quality of non-randomized studies, especially for the purposes of meta-analysis. Nevertheless, as with randomized trials[14] for which there is no gold standard, it is possible that any newly proposed instrument might have internal flaws. An ideal index should be highly sensitive (by increasing the number of items) and applicable in daily practice (by minimizing user burden). This remains the challenge for epidemiologists and research in this field is in its infancy.

MINORS in our opinion has two important attributes. First, its simplicity in comprising only 12 items that are readily usable by both readers and researchers and second, its reliability, as demonstrated by clinimetric testing. Our aim now is to use MINORS in several more studies designed to evaluate the methodology of non-randomized studies. Only the repeated use of such an instrument can confirm the present preliminary clinimetric validation.

ACKNOWLEDGEMENTS

We thank the following participants for their help in the phases of conceptualization and validation: L. Audigé (from the Non-randomized Study Method Group of the Cochrane Collaboration), E. Albuisson, J. E. Bazin, F. Blanchard, G. Borges Da Silva, S. Bouee, D. Chopin, B. Dousset, C. Dziri, F. Fagnani, P-L Fagniez, L. Gerbaud, A. Kramar, F. Lacaine, B. Millat, E. Monnet, P. Perez, A. Perillat, T. Perniceni, L. R. Salmi and the surgeons who returned the completed form during the conceptualization phase of MINORS.

REFERENCES

1. Slim K, Haugh M, Fagniez P-L, Pezet D, Chipponi J. Ten-year audit of randomised trials in digestive surgery from Europe. *Br. J. Surg.* 2000; **87**: 1585–6.
2. McCulloch PM, Taylor I, Sasako M, Lovett B, Griffin D. Randomised trials in surgery: problems and possible solutions. *BMJ* 2002; **324**: 1448–51.
3. Stroup DF, Berlin JA, Morton SC *et al.* Meta-analysis of observational studies in epidemiology. A proposal for reporting. *JAMA* 2000; **283**: 2008–12.
4. Bhandari M, Morrow F, Kulkarni AV, Tornetta P. Meta-analyses in orthopaedic surgery. A systematic review of their methodologies. *J. Bone Joint Surg. Am.* 2001; **83**: 15–24.
5. Khan KS, ter Riet G, Popay J, Nixon J, Kleijnen J. Study quality assessment. In: *Undertaking systematic reviews of research on effectiveness. CRD's guidance for those carrying out or commissioning reviews* CRD Report 4, 2nd edn. York: York Publishing Services; March 2001.
6. Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *J. Epidemiol. Community Health* 1998; **52**: 377–84.
7. Oxman AD, Guyatt GH. Validation of an index of the quality of review articles. *J. Clin. Epidemiol.* 1991; **44**: 1271–8.
8. Feinstein AR. *Clinimetrics.* New Haven: Yale University Press, 1987.
9. Moher D, Schulz KF, Altman DG for the CONSORT Group. CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. *Lancet* 2001; **357**: 1191–4.
10. King TS, Chinchilli VM. A generalized concordance correlation coefficient for continuous and categorical data. *Stat. Med.* 2001; **20**: 2131–47.
11. Bland JM, Altman DG. Cronbach's alpha. *BMJ* 1997; **314**: 572.
12. Bland JM, Altman DG. Validating scales and indexes. *BMJ* 2002; **324**: 606–7.
13. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995; **273**: 408–12.
14. Jüni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA* 1999; **282**: 1054–60.