Int. J. Morphol.,
36(2):762-767, 2018.

Initial Validation of a Scale to Measure Methodological Quality in Prognosis Studies. The MInCir Proposal

Validación Inicial de una Escala para Medir Calidad Metodológica en Estudios de Pronóstico. La Propuesta de MInCir

Carlos Manterola[1,2]; Daniela Zavando[3]; Ricardo Cartes-Velásquez[4]; Tamara Otzen[1,5]; Antonio Sanhueza[6] & MInCir Group (Methodology for Research in Surgery)

MANTEROLA, C.; ZAVANDO, D.; CARTES-VELÁSQUEZ, R.; OTZEN, T.; SANHUEZA, A. & MINCIR GROUP (METHODOLOGY FOR RESEARCH IN SURGERY). Initial validation of a scale to measure methodological quality in prognosis studies. The MInCir proposal. *Int. J. Morphol., 36(2)*:762-767, 2018.

SUMMARY: Research in methodological quality (MQ) of prognosis studies (PS) is relevant in view of the important number of studies developed in this scenario. However, currently there are no instruments designed to measure MQ in PS, thus the aim of this study was to validate a scale to determine the MQ in PS. Scale validation study. Two independent researchers applied the scale (10 items/4 domains) in 119 articles found in 13 Journals of high, medium and low impact factor. Criterion validity was determined by contrasting MQ scores with Oxford Centre for Evidence-Based Medicine levels of evidence. Construct validity of extreme groups and high and low impact factors were estimated. Intraclass correlation coefficient was used to determine interobserver reliability, and the cut-off point was calculated using a ROC curve. The best cut-off point was 33, with an under curve area of 82.6 %. Criterion and construct validity were statistically significant with (p<0.001). Interobserver reliability was 0.91 and a scale to measure the MQ in PS was validated.

KEY WORDS: "Prognosis"[MeSH]; "Validation Studies"[Publication Type]; "Reproducibility of Results"[MeSH]; "Weights and Measures"[MeSH]; Methodological Quality; "Evidence-Based Medicine"[MeSH].

INTRODUCTION

The measurement process is an inherent component of scientific research in any of its disciplines. This allows defining dimensions and generating classifications, thereby facilitating description and communication of the results, which is a daily responsibility for clinicians and researchers (Streiner & Geoffrey, 2003).

We continually estimate the quality of scientific articles for which instruments have been developed, to improve research reading and communication. Specifically, these instruments should allow the user to carefully analyse the results of healthcare research, and determine potential bias in the execution of these studies (Hirst & Altman, 2012; Stevens *et al*., 2014). However, within the range of existing tools this objective appears to be quite different and, rather than assessing methodological quality (MQ), is geared toward improving the quality of reporting results (scoring systems or "checklists", e.g. CONSORT, STARD, STROBE, QUORUM and TREND (Hirst & Altman; Manterola *et al*.).

The construct of MQ is multidimensional (Manterola *et al*., 2006). Although there is no globally accepted version of its components (Armijo-Olivo *et al*., 2013), it can be interpreted as a complex and multidimensional construct, that may include items and domains, such as, type of design, sample size, methodology, analysis quality and reporting quality. The above can be represented as a geometric figure of as many sides as there are domains incorporated into the construct (Fig. 1).

The MQ assessment is an essential step in increasing internal and external validity in research, which would influence the quality of articles published in journals (Manterola *et al*., 2006). Instruments such as QUADAS,

[1] Center of Morphological and Surgical Studies (CEMyQ), Universidad de La Frontera, Temuco, Chile.
[2] Department of Surgery, Universidad de La Frontera, Temuco, Chile.
[3] Municipal Health Department, Machalí, Machalí, Chile.
[4] School of Dentistry, Universidad de Concepción, Concepción, Chile.
[5] Faculty of Health Sciences, Universidad de Tarapacá, Arica, Chile.
[6] Pan-American Health Organization, Washington, USA.

Table I. MQ scale for PS.

Domains and items of the scale	**Score (points)**
Domain 1: Study design (type of study).	
Validating studies test with good reference standards *	**12**
Exploratory cohort study with reference standards **	**9**
Validating studies test with poor reference standards ***	**6**
Case control study poor or non-independent reference standard	**4**
Poor quality cohort studies ****	**3**
Case-series	**1**
Domain 2: Sample size x justification factor (duplication of original value).	
≥ 201	**6** or **12**
151–200	**5** or **10**
101–150	**4** or **8**
61–100	**3** or **6**
31–60	**2** or **4**
≤ 30	**1** or **2**
Domain 3: Methodology.	
Item 1. Objectives	
Clear and concrete objectives	**3**
Vague objectives	**2**
No objective	**1**
Item 2. Design	
Clearly identified design	**3**
Unknown design	**1**
Item 3. Selection criteria	
Inclusion and exclusion criteria are described	**3**
Inclusion or exclusion is described	**2**
No selection criteria are described	**1**
Item 4. Characterisation of the population under study	
The spectrum of the study subjects is representative of the population for which it is desired to extrapolate the results.	**3**
The spectrum of the study subjects is partially representative of the population for which it is desired to extrapolate the results	**1**
Item 5. Characteristics of the reference standard applied	
The same reference standard is applied to all the patients independent of the results	**3**
The reference standard is applied partially or differentially	**2**
There is no report of the standard of reference used	**1**
Item 6. Characteristics of the diagnostic test under study	
The diagnostic test under study is described with sufficient detail to allow the replication.	**3**
The diagnostic test under study is only partially described.	**2**
The authors do not provide elements concerning the diagnostic test under study that allow for the study to be replicated.	**1**
Item 7. Sample size	
Sample size is justified	**3**
Sample size is not justified	**1**
Final score _ (domains 1 + 2 + 3)	**9-45**

**: Studies that justify the sample sizes have double scores.

AMSTAR and others (Shea *et al*., 2007; Mokkink *et al*., 2009; Manterola *et al*., 2013), have been used for these purposes.

Although, there are at least 26 validated instruments for measuring the MQ of randomised clinical trials (Armijo-Olivo *et al*.), there is no single tool to determine MQ in prognosis studies (PS). There is only a proposal to evaluate PS quality in systematic reviews (SR) (Hayden *et al*., 2006), a fact that hinders the development of SR in the clinical research scenario.

The recently created MInCir scale (Methodology for Research in Surgery) to assess MQ in PS, is composed of 10 items grouped into 4 domains (Table I). Instructions for its use, have been published with the aim of providing a guideline for its standardized application (Manterola *et al*., 2015).

The aim of this study was to validate a scale to determine the MQ in PS.

MATERIAL AND METHOD

Study design: Scale validation study (Streiner & Geoffrey).

Setting: Center of Excellence in Morphological and Surgical Studies, and PhD Program in Medical Sciences. Universidad de La Frontera, Chile.

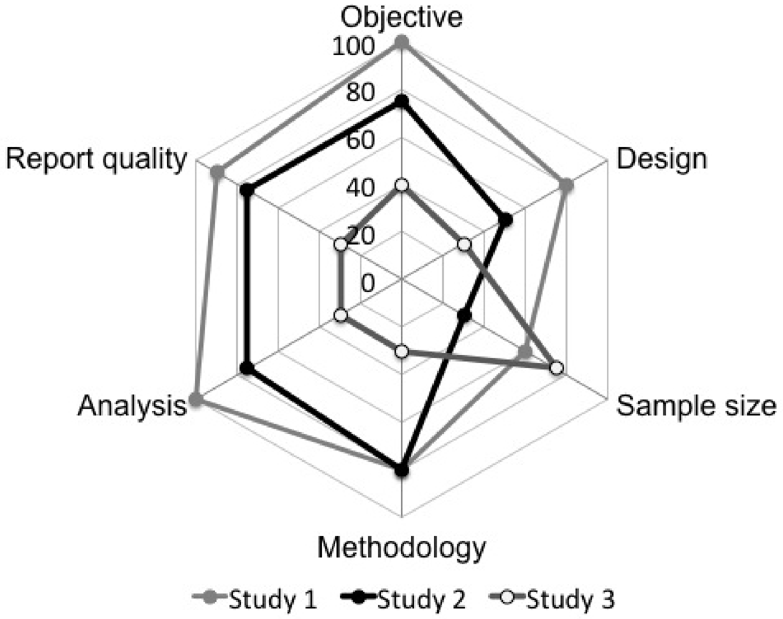

Fig. 1. Polar graph in which 6 domains are measured to explain the MQ construct. Three primary articles that occupy different surfaces of the hexagon may be appreciated. Number one indicates the best performance of the construct. Data included in this graph are three "examples".

Validation study: Primary articles of prognosis published in 13 journals in Spanish and English were grouped according to their impact factor (IF): high [≥4], medium [3.99-0.5] and low [≤0.499] (Thomson Reuters, 2018). A simple random sample of 140 articles was performed applying Streiner feasibility criteria, there should be at least 10 articles per item (Streiner & Geoffrey). Of the random sample (140), 21 did not allow for the evaluation of instrument items due to problems with the quality of previously reported results. Two researchers (DZ and CM) independently applied the instrument in a sample of 119 articles, settling disagreements by consensus, thereby obtaining two independent scores and one consensual score.

Criterion and construct validity: Using the consensus score, criterion validity was determined by contrasting MQ scores with the levels of evidence of the Oxford Centre for Evidence-Based Medicine (CEBM, 2009). Construct validity was determined via extreme group analysis by dichotomising the IF of journals in which the aforementioned studies were published.

Interobserver reliability: Using the two independent scores, the degree of agreement between evaluators was determined.

Statistical analysis: Measures of central tendency and dispersion were used. The cut-off point was determined using ROC curve. Criterion validity was determined using one-factor ANOVA and Duncan test. Construct validity was determined applying T-Test, and interobserver reliabilityby applying an intraclass correlation coefficient. Analyses were performed using STATA 10/SE (Stata Corp., TX, USA).

RESULTS

For sensitivity and specificity analysis, area under ROC curve was 82.6 % (Fig. 2). The best cut-off point was 33 to define the MQ construct and differentiate a good and poor MQ in PS (Table II).

Criterion validity was determined comparing the MQ with the levels of evidence. The levels of evidence of the sample of articles were 1b (21 articles, 17.6 %), 2b (15 articles, 12.6 %) and 4 (83 articles, 69.8 %), which had mean MQ scores of 45.3±3.0, 40.8±2.8 and 27.3±2.5, respectively (p<0.0001), demonstrating the criterion validity of the scale.

Articles with high and low IF obtained mean MQ scores of 38.8±8.0 and 27.6±7.6, respectively (p<0.0001); criterion validity was checked in this way. Intraclass correlation coefficient for interobserver reliability was 0.91 (p=0.1082). In 52.9 % of the sample MQ was good. A description of the distribution of the total scores disaggregated by domain is presented in Table III.

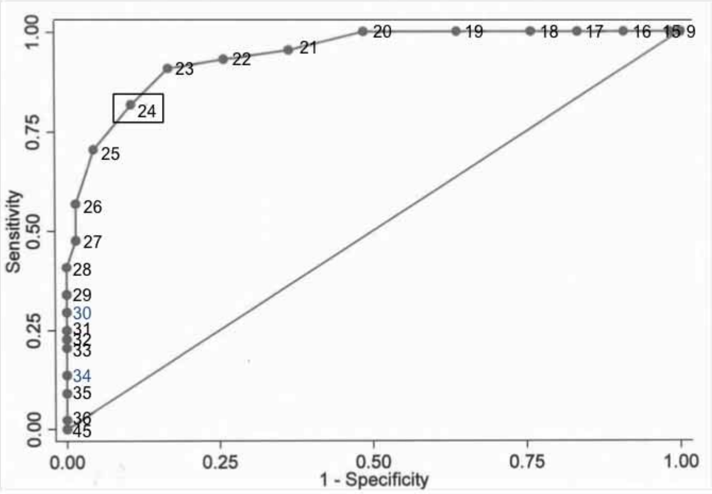

Fig. 2. ROC curve that shows the cut-off point that defines the MQ construct (33 points) and the area under the curve.

Table II. Analysis of psychometric parameters and the cut-off points representative of the scale.

Parameters	Cut-off points			
	22	23	24	25
Sensitivity (%)	84.1	81.8	79.6	70.5
Specificity (%)	84.9	90.9	93.9	97.0
PPV (%)	78.7	85.7	89.7	94.0
NPV (%)	88.9	88.2	87.3	83.1
LHR (+)	5.6	9.0	13.1	23.3
LHR (-)	0.19	0.20	0.22	0.30
Correct classification (%)	84.6	87.3	88.2	86.4
Area under curve (%)	85.0	86.0	87.0	84.0
Association article/scale	29.6	45.0	60.0	76.3

OR :Odds Ratio

Table III. Distribution of total score by domain of the scale of MQ in PS.

Statistics	Domain scores			Total score
	1	2	3	
Average ± SD	3.1±3.3	4.5±2.4	14.6±2.6	22.1±6.6
Median	1	4	15	21
Interquartile range	1-4	3-6	12-16	18-25
Minimum and maximum *	1-12	1-12	9-21	13-45

SD : Standard Deviation

DISCUSSION

To determine instrument´s psychometric properties, a number of working strategies are necessary (Rojahn *et al*., 2011). Therefore, the aim was to provide a useful tool to determine MQ of PS, because current instruments are scarce and their quality is questionable (Hayden *et al*., 2006; Armijo-Olivo *et al*.).

We identified items and domains to define the construct of MQ in PS by generating a scale and assessing its performance in a sample of articles from different biomedical journals.

The scale design was developed in three steps: First, an item selection to generate the first draft (this step was carried out by a review of the literature of MQ of PS, and conducted via a systematic search in BIREME, PubMed, OVIDWeb, Scopus, Web of Science and SciELO. An expert panel of five clinical epidemiologists and one biostatistician suggested the items and domains from which to build the construct (this was based on the literature review and their personal experience). In order to generate a second draft an alphanumeric order was given, and the draft was evaluated by researchers from USA, Spain and Chile, with Master's or Ph.D. degrees in Medical Sciences, and at least one publication related to MQ, in the Web of Science database. A pilot study allowed a third draft that involved graduate students of the Ph.D. and Master's programs in medical sciences, of the Universidad de La Frontera. It optimized the use and understanding of the scale, and the third draft comprising four domains and 12 items, was obtained.

Criterion validity was demonstrated via contrasting the scores obtained by applying the scale to the sample of articles, with the levels of evidence of each one. However, this estimate should be viewed with caution, since the levels of evidence are intended to value only the design of the studies (Manterola *et al*., 2013), which is only one dimension evaluated when using this MQ construct. However, even if the level of evidence is an imperfect standard for MQ, it is useful to clearly distinguish prognosis scenarios (CEBM).

On the other hand, construct validity of extreme groups, is based on the IF of the journals that published the articles used in the analysis. The criterion seems reasonable, given the reported correlation between the IF for medical journals and the MQ of each article, as perceived by clinicians and biomedical researchers (Saha *et al*., 2003; Manterola *et al*., 2005a). However, there is evidence of the controversial nature of using the IF as an indicator of MQ (Manterola *et al*., 2006), since high-IF journals publish studies of low MQ (Gluud *et al*., 2005). There is also information that dismisses the IF as an adequate measure to assess MQ (Favaloro, 2008). The use of articles in different languages (English and Spanish) could be another source of bias to the limitations already mentioned with respect to the IF.

The high interobserver reliability could possibly be the result of the examiners' common views of the MQ construct and the interpretation of the scale under discussion. Hence the importance of calibrating observers to obtain reproducible results (Manterola *et al*., 2015).

The scale, allows us not only to analyse an article based on a question of clinical use, but also to conduct bibliometric studies as in therapy scenarios (Manterola *et al*., 2006). SR and meta-analyses of information can also be carried out by weighting the available evidence (Manterola *et al*., 2009). This instrument therefore, represents a critical contribution to clinical research.

There is another tool for assessing MQ in PS (Quality in Prognosis Studies (QUIPS), which has recently been updated (Hayden *et al*., 2013). This tool evaluates the quality of PS in SR. However, the approach is different in that it focuses on identifying research biases. It also presents differences in its development (Hayden *et al*., 2006, 2008, 2013), which relate to the continuous refinement of the instrument by experts who recommend using it to evaluate bias in its six domains (Hayden *et al*., 2006, 2008, 2013).

This new scale will allow determining the MQ in PS, facilitating the work of readers, authors, reviewers and editors of biomedical journals. At the same time, it can be used to conduct bibliometric studies, and for weighting the quality of evidence through meta-analyses of information obtained from SR in prognosis scenarios, with primary studies of different types of designs.

Notwithstanding the above, the methods of this type of studies are not clearly defined. The scale uses specific items; the decision to classify the articles into three categories based on the IF distribution of WoS database journals; the cut-off points to discriminate the MQ dichotomously as good and bad; the choice to use the cut-off point 33 (based in the best under curve area (82.6 %) and correct classification (75.6 %), in addition to having good sensitivity (80.0 %), positive predictive value (82.6 %), positive likelihood ratio (3.01) and odds ratio (12.2).

Our purpose is to continue developing this scale and publish new information. We believe that the next steps should include confirming the psychometric properties of the scale in different specialties and disciplines, analyzing the correlation with QUIPS, and performing bibliometric studies and SR with PS. Thus, we believe that this scale can be used to define the MQ in PS, but we do not want it to become a static instrument.

In conclusion, we can point out that a scale to measure the MQ in PS was validated.

MANTEROLA, C.; ZAVANDO, D.; CARTES-VELÁSQUEZ, R.; OTZEN, T.; SANHUEZA, A. & MINCIR GROUP (METHODOLOGY FOR RESEARCH IN SURGERY). Validación inicial de una escala para medir calidad metodológica en estudios de pronóstico. La propuesta de MInCir. *Int. J. Morphol., 36(2)*:762-767, 2018.

RESUMEN: El objetivo de este estudio fue validar una escala para determinar calidad metodológica (CM) de estudios de pronóstico (EP). Se realizó un estudio de validación de escalas. La escala, compuesta por 10 ítems y 4 dominios; se aplicó a 119 artículos de 13 revistas, de factores de impacto alto, medio y bajo; por dos investigadores independientes. La validez del criterio se determinó al contrastar las puntuaciones de CM de cada artículo con los niveles de evidencia del Centro de Medicina Basada en la Evidencia de Oxford de la revista en la cual fueron publicados. Se estimó la validez de constructo de grupos extremos (factores de impacto alto y bajo). Se utilizó el coeficiente de correlación intraclase para determinar la confiabilidad interobservador, y el punto de corte se calculó construyendo curvas ROC. El mejor punto de corte fue 33 puntos (área bajo la curva de 82,6 %). La validez de criterio y de constructo fueron estadísticamente significativas (p<0,001). La confiabilidad interobservador fue 0,91. Se validó una escala para medir CM en EP.

PALABRAS CLAVE: Calidad metodológica; Pronóstico; Estudios de validación de escalas; Medición; Medicina basada en evidencia.

REFERENCES

Armijo-Olivo, S.; Fuentes, J.; Ospina, M.; Saltaji, H. & Hartling, L. Inconsistency in the items included in tools used in general health research and physical therapy to evaluate the methodological quality of randomized controlled trials: a descriptive analysis. *B. M. C. Med. Res. Methodol., 13*:116, 2013.

Centre for Evidence-Based Medicine (CEBM). *Oxford Centre for Evidence-based Medicine - Levels of Evidence (March 2009)*. Oxford, Centre for Evidence-Based Medicine (CEBM), 2009. Available from: http://www.cebm.net/oxford-centre-evidence-based-medicine-levels-evidence-march-2009

Favaloro, E. J. Measuring the quality of journals and journal articles: the impact factor tells but a portion of the story. *Semin. Thromb. Hemost., 34(1)*:7-25, 2008.

Gluud, L. L.; Sørensen, T. I.; Gøtzsche, P. C. & Gluud, C. The journal impact factor as a predictor of trial quality and outcomes: cohort study of hepatobiliary randomized clinical trials. *Am. J. Gastroenterol., 100(11)*:2431-5, 2005.

Hayden, J. A.; Côté, P. & Bombardier, C. Evaluation of the quality of prognosis studies in systematic reviews. *Ann. Intern. Med., 144(6)*:427-37, 2006.

Hayden, J. A.; Côté, P.; Steenstra, I. A.; Bombardier, C. & QUIPS-LBP Working Group. Identifying phases of investigation helps planning, appraising, and applying the results of explanatory prognosis studies. *J. Clin. Epidemiol., 61(6)*:552-60, 2008.

Hayden, J. A.; van der Windt, D. A.; Cartwright, J. L.; Côté, P. & Bombardier, C. Assessing bias in studies of prognostic factors. *Ann. Intern. Med., 158(4)*:280-6, 2013.

Hirst, A. & Altman, D. G. Are peer reviewers encouraged to use reporting guidelines? A survey of 116 health research journals. *PLoS One, 7(4)*:e35621, 2012.

Manterola, C.; Cartes-Velásquez, R. & Otzen, T. Instructions for using the MInCir scale to assess methodological quality in diagnostic accuracy studies. *Int. J. Morphol., 34(1)*:78-84, 2016.

Manterola, C.; Otzen, T.; Lorenzini, N.; Díaz, A.; Torres-Quevedo, R. & Claros, N. Initiatives for reporting biomedical research results with different types of designs. *Int. J. Morphol., 31(3)*:945-56, 2013.

Manterola, C.; Pineda, V. & Vial, M. Open versus laparoscopic resection in non-complicated colon cancer. A systematic review. *Cir. Esp., 78(1)*:28-33, 2005b.

Manterola, C.; Pineda, V.; Vial, M. & Losada, H. Is impact factor an appropriate index to determine the level of evidence of studies on therapeutic procedures in surgery journals? *Cir. Esp., 78(2)*:96-9, 2005a.

Manterola, C.; Pineda, V.; Vial, M.; Losada, H. & MINCIR Group. What is the methodologic quality of human therapy studies in ISI surgical publications? *Ann. Surg., 244(5)*:827-32, 2006.

Manterola, C.; Pineda, V.; Vial, M.; Losada, H. & Muñoz, S. Surgery for morbid obesity: selection of operation based on evidence from literature review. *Obes. Surg., 15(1)*:106-13, 2005c.

Manterola, C.; Vial, M.; Pineda, V. & Sanhueza, A. Systematic review of literature with different types of designs. *Int. J. Morphol., 27(4)*:1179-86, 2009.

Mokkink, L. B.; Terwee, C. B.; Stratford, P. W.; Alonso, J.; Patrick, D. L.; Riphagen, I.; Knol, D. L.; Bouter, L. M. & de Vet, H. C. Evaluation of the methodological quality of systematic reviews of health status measurement instruments. *Qual. Life Res., 18(3)*:313-33, 2009.

Rojahn, J.; Rowe, E. W.; Kasdan, S.; Moore, L. & van Ingen, D. J. Psychometric properties of the Aberrant Behavior Checklist, the Anxiety, Depression and Mood Scale, the Assessment of Dual Diagnosis and the Social Performance Survey Schedule in adults with intellectual disabilities. *Res. Dev. Disabil., 32(6)*:2309-20, 2011.

Saha, S.; Saint, S. & Christakis, D. A. Impact factor: a valid measure of journal quality? *J. Med. Libr. Assoc., 91(1)*:42-6, 2003.

Shea, B. J.; Grimshaw, J. M.; Wells, G. A.; Boers, M.; Andersson, N.; Hamel, C.; Porter, A. C.; Tugwell, P.; Moher, D. & Bouter, L. M. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. *B. M. C. Med. Res. Methodol., 7*:10, 2007.

Stevens, A.; Shamseer, L.; Weinstein, E.; Yazdi, F.; Turner, L.; Thielman, J.; Altman, D. G.; Hirst, A.; Hoey, J.; Palepu, A.; Schulz, K. F. & Moher, D. Relation of completeness of reporting of health research to journals' endorsement of reporting guidelines: systematic review. *B. M. J., 348*:g3804, 2014.

Streiner, D. L. N. & Geoffrey, R. *Health Measurement Scales: A Practical Guide to Their Development and Use*. 3rd ed. Cary, Oxford University Press, 2003.

Thomson Reuters. *JCR-Web 4.5 Category Selection*. 2018. Available from: http://admin-apps.webofknowledge.com.ezproxy.puc.cl/JCR/JCR

Corresponding author:
Prof. Dr. Carlos Manterola
Center of Morphological and Surgical Studies (CEMyQ)
Universidad de La Frontera
Avenida Francisco Salazar 01145
Temuco
CHILE

E-mail: carlos.manterola@ufrontera.cl

Received: 04-01-2018
Accepted: 18-03-2018