Int. J. Morphol.,
36(2):743-749, 2018.

Development and Initial Validation of a Scale to Measure Methodological Quality in Diagnostic Accuracy Studies. The MInCir Proposal

Desarrollo y Validación Inicial de una Escala para Medir la Calidad Metodológica en Estudios de Precisión Diagnóstica. La Propuesta de MInCir.

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

MANTEROLA, C.; CARTES-VELÁSQUEZ, R.; BURGOS, M. E.; SANHUEZA, A.; OTZEN, T. & MINCIR GROUP (METHODOLOGY FOR RESEARCH IN SURGERY). Development and initial validation of a scale to measure methodological quality in diagnostic accuracy studies. The MInCir proposal. *Int. J. Morphol., 36(2)*:743-749, 2018.

SUMMARY: Research in diagnostic accuracy studies (DAS) is a rapidly developing area in medicine, but there are only three instruments used in this scenario. The aim of this study was to design and validate a scale to determine methodological quality (MQ) of DAS. Scale validation study. A systematic literature review about the MQ of diagnostic accuracy studies was accomplished, and an expert panel generated a first draft (content validity) of the scale. An alphanumeric order was given and rated by six researchers (second draft) and a pilot study to optimise its use and understanding was performed (third draft). Two independent researchers applied the final scale (9 items/3 domains) to 110 articles from 13 journals with high, medium and low impact factors. Criterion validity was determined by contrasting MQ scores with the Oxford Centre for Evidence-Based Medicine levels of evidence. The construct validity of the extreme groups and high and low IF were estimated. The intraclass correlation coefficient was used to determine inter-observer reliability, and the cut-off point was calculated using a ROC curve. The best cut-off point was 24 points, with an under curve area of 93.4 %. The content validity rating was 80–100 % for all included items. Criterion and construct validity were statistically significant with $p<0.05$. Interobserver reliability was estimated in 0.96. A scale to measure the MQ of DAS was designed and validated.

KEY WORDS: Diagnostic Techniques and Procedures; Validation Studies[Publication Type]; Reproducibility of Results; Weights and Measures; Evidence-Based Medicine.

INTRODUCTION

With the exponential growth of scientific information, it is difficult to cover everything that is published. On the other hand, not every article has the same value. from the point of view of Evidence-Based Medicine. Therefore, researchers and clinicians need to acquire the competence to critically appraise the evidence, identifying good quality studies and ensuring optimal patient care (du Prel *et al.*, 2009).

One of the key aspects to be evaluated in a scientific article is the methodological quality (MQ), and the process is complex because the MQ construct is multidimensional. It is possible to evaluate multiple items and domains such as: design, sample size, methodology, analysis quality, reporting quality, etc. All of these dimensions can be represented in a geometric figure of, as many sides as domains are incorporated in the construct (Manterola *et al.*, 2006) (Fig. 1). However, this construct does not currently have a single definition (Armijo-Olivo *et al.*, 2013), and many different instruments have been developed to recognize research biases and the applicability of the results in clinical practice (Mokkink *et al.*, 2009).

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

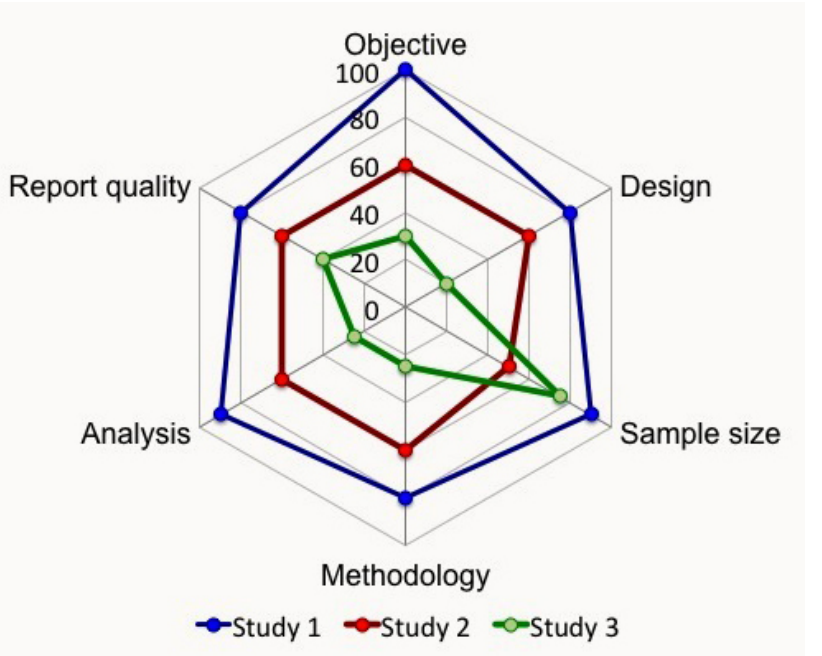

Fig. 1. Polar graph in which six domains are measured to explain the MQ construct. Three primary articles that occupy different surfaces of the hexagon may be appreciated. Number one represents a study of good MQ while number three represents a study of poor MQ.

Research in diagnostic accuracy studies (DAS) is a rapidly developing area in medicine, but there are only three international instruments used in this scenario (Oliveira *et al*., 2001; Saha *et al*., 2003; Bossuyt *et al*., 2003; Whiting *et al*., 2006; Mokkink *et al*., 2009; Manterola *et al*., 2013; Cook *et al*., 2014). STARD is a reporting guideline with a checklist (Oliveira *et al*., 2001; Bossuyt *et al*.; Manterola *et al*., 2013; Cook *et al*.); QUADAS and QUADAS-2 assess different aspects of quality across 4 domains, and use signalling questions to tailor assessments (QUADAS-2, probably the most commonly used quality assessment tool for DAS, was developed to deal with some of the inadequacies of the original QUADAS tool) (Oliveira *et al*., 2001; Whiting *et al*., 2003, 2006, 2011; Cook *et al*.). QUADAS was a checklist that included components of underlying methodological quality, as well as quality of reporting. This was identified as being problematic and was corrected in QUADAS-2, which focuses only on the components of methodological quality that may lead to bias, and on applicability (external validity). It was intentionally developed to not include a numerical assessment, score or scale, as there is evidence that score/scale based quality assessment tools for any study design, are trivial and meaningless (Oliveira *et al*., 2001; Manterola *et al*., 2013).

On the other hand, MQ assessment is a crucial step to increase the internal and external validity of articles, influencing the quality of journals (Manterola *et al*., 2006).

Our working group has developed scales to measure MQ in different scenarios, with the purpose of performing systematic reviews (SR) with different designs, in addition to bibliometric studies (Manterola *et al*., 2009). The MInCir scale (Metodología de Investigación en Cirugía/ Methodology for Research in Surgery) for assessing the MQ in DAS was recently developed, and the instructions for its use have been published with the aim of providing a guideline for its standardized use (Manterola *et al*., 2016).

The aim of this study was to design and validate a scale for determine the MQ of DAS.

MATERIAL AND METHOD

Study design: Scale validation study (Streiner & Geoffrey, 2003).

Setting: Center of Morphological and Surgical Studies and Department of Surgery. Universidad de La Frontera, Chile. Scale design:

Item selection (first draft): A review of the literature about MQ of DAS was conducted via systematic search in libraries and databases BIREME, PubMed, OVIDWeb, Scopus, Web

of Science and SciELO with the following strategy: (("methodological studies" OR "validation studies") AND "accuracy" AND "diagnostic"). All SR of level 1, 2 and 3 diagnostic studies, validating cohort studies with good reference standards, exploratory cohort studies with good reference standards, non-consecutive studies or without consistently applied reference standards, case-control studies, poor or non-independent reference standard, in human population, published in the last 5 years in the English language were included (N = 654). Then, through application of Delphi method to refine an initial list of items, an expert panel comprised of five clinical epidemiologists and one biostatistician suggested the items and domains from which to build the construct of MQ for DAS, based on the literature review and their personal experience in MQ.

Content validation (second draft): Content validity was defined as the extent to which a measure represents all facets of a given construct. It requires the use of recognized subject matter experts to evaluate whether test items assess defined content (Wilson *et al*., 2012).

An alphanumeric order was given and the second draft was created. Five researchers (one from the USA, one from Spain and three from Chile) evaluated this draft. All of them had experience in the field (with master's or doctoral degrees in medical sciences, with at least one publication in the Web of Science database related to MQ). The experts assessed the relevance of each item with a 1–7 Likert scale and provided comments to improve the instrument.

Pilot study (third draft): A pilot study involving graduate students in the field of medical sciences was conducted (three from Ph.D. and three from master's degree programs). This was performed in order to optimize the use and understanding of the scale. Using a Likert scale, with the possibility of making comments, this assessment was also objectivized. Thus, the third draft was comprised of three domains and nine items, with a minimum of 9 and a maximum of 45 points (Table I).

Validation study: A simple random sample of 110 primary articles of diagnosis accuracy was included. Inclusion criteria were cross-sectional and case-control studies in humans, with no limits of language, age or year of publication. The articles were published in 13 journals in Spanish and English, and were grouped according to their impact factor (IF) in: high [≥3], medium [3 to 1] and low [<1] (Thomson Reuters, 2018). Subsequently, two researchers (CM and MB) independently applied the instrument to the sample of articles, settling disagreements by consensus; with this information they obtained two independent scores and one consensus score.

Criterion 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). Levels of evidence were used as an ordinal variable, categorized from 1 (evidence level 1) to 4 (evidence level 4).

Construct validity: Construct validity was determined through extreme group analysis by dichotomising the IF of the journals in which the aforementioned articles were published and, assuming that high-IF journals publish articles of better MQ.

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

Statistical analysis: Measures of central tendency and dispersion were used (average and standard deviation). Internal consistency was estimated using Cronbach's alpha. The cut-off point was determined using the Receiver Operating Characteristic (ROC) curve and the criterion validity was determined using the Spearman correlation. Construct validity was calculated by applying linear regression, and inter-observer reliability was determined by applying the intraclass correlation coefficient. All analyses were made using STATA 10/SE (Stata Corp., TX, USA).

RESULTS

The mean IF of the sample of 13 journals included in the study was 4.3±8.2. Content validity, according to the expert opinion, was between 80 % and 100 % among all items included.

The internal consistency was estimated at 0.60.The area under the ROC curve was 93.4 % (Fig. 2).

The analysis of diagnostic parameters of the instrument determined a cut-off point of 24 to define the MQ construct and differentiate between good and poor MQ for DAS (Table II).

Levels of evidence of the sample of articles showed a high correlation (0.79), with the scale score ($p<0.001$), which was used to check the criterion validity. Articles of high and low IF received mean scores of MQ of 25.4±8.9 and 19.8±5.5, respectively ($p=0.03$), which was used to check construct validity.

Table I. MQ scale for DAS.

Domains and items of the scale	Score
Domain 1: Research design	
Concurrent or prospective cohort. Controlled, double-blind, randomised, clinical trial	15
Historical or retrospective cohort. Non-randomised clinical trial	10
Case control study	8
Cross-sectional study	6
Case report or case series	3
Domain 2: Studied population x justification factor**	
> 501	7 or 15
201–500	6 or 12
151–200	5 or 10
101–150	4 or 8
51–100	3 or 6
31–50	2 or 4
≤ 30	1 or 2
Domain 3: Methodology	
Objective	
Clear and concrete objectives	3
Vague objectives	2
No objectives	1
Design	
Clearly identified the design	3
Unknown design	1
Variables (definition of outcome, exposure and confounding variables)	
Outcome variables adequately defined	1 or 0
Exposure variables adequately defined	1 or 0
Confounding variables adequately defined	1 or 0
Sample size	
Includes sample size calculation/estimation	3
Does not include sample size calculation/estimation	1
Follow-up	
Mentioned the losses/follow-up percentage	1 or 0
The follow-up was greater than 80%	1 or 0
Cause of losses explained	1 or 0
Domain 4: Analysis and conclusions	
Risk measures	
Included a calculation of the risk measures	5 or 0
Reported data allowed the calculation of risk measures	2 or 0
Association models	
Included predictive or association models	5 or 0
Consistency between objective, methodology and results	
Showed consistent objective-methodology-results	3 or 0
Total _ (domains 1 + 2 + 3 + 4)	**7–60**

* : Validating studies test the quality of a specific diagnostic test, based on prior evidence. Good reference standards are independent of the test, and applied blindly or applied objectively to all patients.
**: Exploratory studies collect evidence and search the data to find which factors are important.
***: Poor reference standards are arbitrarily applied, but independent of the test.
****: Includes non-consecutive study without consistently applied reference standards.

The intraclass correlation coefficient for inter-observer reliability was 0.96. A description of the distribution of total scores and the distribution of the scores disaggregated by domains are presented in Table III.

Table II. Psychometric parameters of different cut-off points of the scale.

Parameters	Cut-off points			
	31	**32**	**33**	**34**
Sensitivity (%)	85.0	82.5	**80.0**	75.0
Specificity (%)	64.6	69.6	**73.4**	77.2
Positive predictive value (%)	79.4	79.5	**82.6**	83.0
Negative predictive value (%)	73.1	73.9	**72.0**	66.7
Likelihood ratio (+)	2.39	2.72	**3.01**	3.29
Likelihood ratio (-)	0.23	0.25	**0.27**	0.32
Correct classification (%)	71.4	73.9	**75.6**	76.5
Area under curve (%)	78.1	80.9	**82.6**	81.9
Association articles/scale (OR)	10.6	10.9	**12.2**	9.8

PPV = Positive predictive value. NPV = Negative predictive value. LHR (+) = Positive likelihood ratio. LHR (-) = Negative likelihood ratio. OR = Odds ratio.

Table III. Distribution of domains scores of the scale.

Statistics	Domain scores				Total score
	1	2	3	4	
Mean ± SD	5.8 ± 4.7	5.3 ± 2.0	12.6 ± 3.5	9.3 ± 1.5	32.8 ± 8.3
Median	3	6	12	9	33
Interquartile range	3–8	3–7	10–15	8–10	26–36
Minimum and maximum	3–15	1–15	3–15	0-15	7–60

SD = Standard deviation. * = Minimum and maximum values found in the study sample.

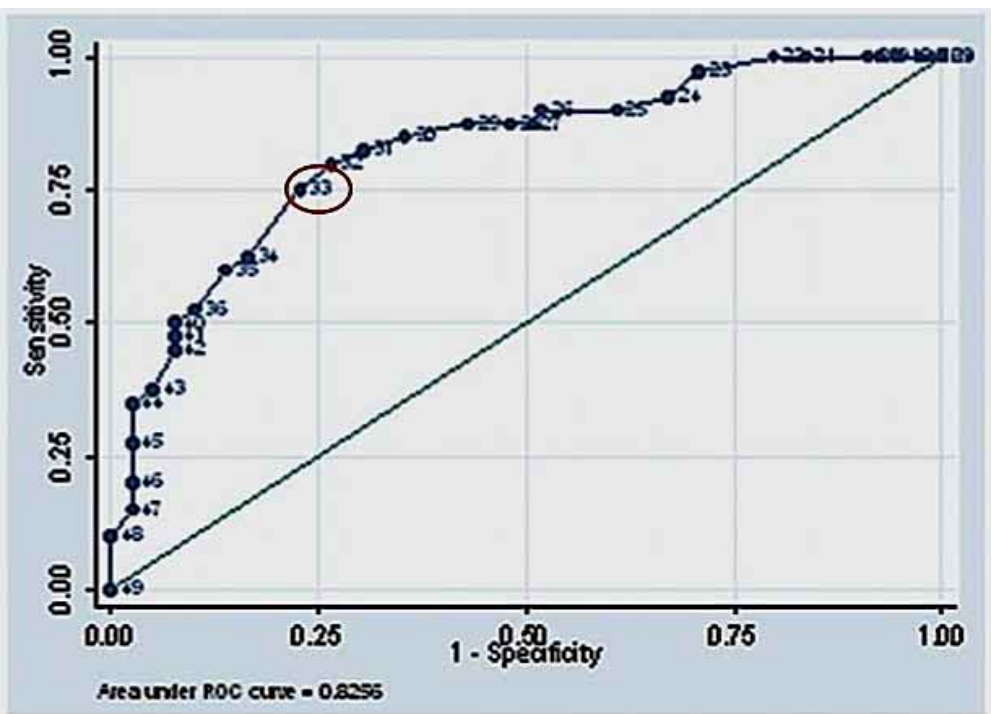

Fig. 2. ROC curve in which the cut-off point that defines MQ construct (24 points) and the area under the curve are seen.

DISCUSSION

In this study, a valid and reliable scale comprised of three domains and nine items for measure the construct MQ for DAS was built and later psychometrically evaluated. MQ of DAS is certainly a problem that affects to the primary studies, because, it has been verified that many studies used in SR of DAS have low MQ (Leeflang *et al*., 2007), which tends to overestimate the capacity of diagnostic tests. Moreover, even when checklists, such as STARD and QUADAS (Oliveira *et al*., 2001), are used, there are few SR where the MQ is considered in their conclusions (Ochodo *et al*., 2014). There are some tools in the DAS scenario, such as STARD and QUADAS (Oliveira *et al*.), and these checklists were therefore used as the basis for the design of our scale. Unfortunately, neither of these instruments has been through a validation process to ensure their psychometric properties (Streiner & Geoffrey).

The internal consistency was acceptable (George & Mallery, 2003), in spite of the inherent difficulties in defining the items and domains of the MQ of DAS (Armijo-Olivo *et al*.) that apply to all types of diagnostic tests, this was already reported in the development of QUADAS (Whiting *et al*., 2003, 2006).

IF was used as a reference standard to determine construct validity by extreme groups, assuming that journals with a higher IF published articles of better MQ (Leeflang *et al*.). However, it must be considered that this is an imperfect standard because high-IF journals also publish articles of low MQ (Favaloro, 2008), so there is a need for alternatives to the current quantitative bibliometric indicators.

Something similar happened with the validity criterion. In this case, the levels of evidence as a reference standard were used. Levels of evidence are an essential aspect of the MQ, but there are other methodological aspects that may differ, in articles with a comparable level of evidence (Oliveira *et al*.; Cartes-Velásquez *et al*., 2014). Despite this limitation, levels of evidence provide a reference standard that is widely accepted in biomedical journals (Joyce *et al*., 2015).

One of the strengths of this scale and others developed by the MInCir group is the high level of inter-observer reliability (Saha *et al*.), which exceeds that of other instruments designed to measure MQ (Armijo-Olivo *et al*.). This is relevant to any instrument that will be used extensively by the research community, and was one of the issues included in improving QUADAS to QUADAS-2 (Whiting *et al*, 2011). To ensure reproducibility of data, it is necessary to have guidelines for using the instruments (Manterola *et al*., 2016) since in many cases, articles are not accurate in this matter (Saha *et al*.) and some items of the scale have some degree of subjectivity (Whiting *et al*., 2011).

The ROC curve analysis allows for defining the 24 points, as the cut-off points for discriminating the MQ dichotomously as good and poor. The choice of these values is always complex and cannot be based only on isolated parameters, such as the area under the curve (Wald & Bestwick, 2014). Notwithstanding the above in this case, the best area under curve (87.0 %) and correct classification (88.2 %), in addition to having good specificity (93.9 %), positive predictive value (89.7 %), positive likelihood ratio (13.1) and odds ratio (60), were the inputs used to define the cut-off point. In this way, a specific use for this instrument would be the assessment of articles with high levels of evidence in order to detect good MQ studies beyond their level of evidence.

Recently, an American group developed the Diagnostic Accuracy Quality Scale (DAQS) (Cook *et al*.) as an alternative to the QUADAS-2 (Whiting *et al*., 2011), as criticism to its use. However, this scale has not yet undergone validation processes as performed in this study. The DAQS has 21 items, which is hardly comparable with the simplicity of the MInCir scale. In addition, the same authors have declared a limitation related to the development of it, because the working group comes mostly from the area of physical therapy. Once the DAQS is validated, comparative studies with our scale could be carried out.

Possible uses of this scale must emphasize the realization of bibliometric studies and SR (Manterola *et al*., 2006), but since this is an initial validation, it is necessary to continue reporting the psychometric properties of the scale in different biomedical disciplines. The MQ is a constantly evolving concept, and therefore, this scale should not be considered a static instrument, but rather should be refined further, as has happened with other instruments in DAS scenarios (Whiting *et al*., 2011).

In conclusion we can point out that a scale to measure the MQ of DAS was designed and validated.

MANTEROLA, C.; CARTES-VELÁSQUEZ, R.; BURGOS, M. E.; SANHUEZA, A.; OTZEN, T. & MINCIR GROUP (METHODOLOGY FOR RESEARCH IN SURGERY). Desarrollo y validación inicial de una escala para medir la calidad metodológica en estudios de precisión diagnóstica. La propuesta de MInCir. *Int. J. Morphol., 36(2)*:743-749, 2018.

RESUMEN: La investigación en estudios de precisión diagnóstica (EPD) es un área de rápido desarrollo en medicina, sin embargo, en este escenario sólo existen tres instrumentos. El objeti-

vo de este estudio fue diseñar y validar una escala para determinar calidad metodológica (CM) de EPD. Estudio de validación de escala. Se realizó una extensa revisión de la literatura sobre el CM de EPD y un panel de expertos generó un primer borrador (validez del contenido) de la escala. Se asignó un orden alfanumérico, el que evaluado por 6 investigadores independientes (2º borrador). Posteriormente, se realizó un estudio piloto para optimizar el uso y entendimiento (3º borrador). Dos investigadores independientes aplicaron la escala final (9 ítems / 3 dominios) a 110 artículos de 13 revistas con factores de impacto alto, medio y bajo. Se determinó validez de criterio contrastando puntuaciones de CM con niveles de evidencia del Oxford Centre for Evidence-Based Medicine. Se determinó validez de constructo de grupos extremos (factores de impacto alto y bajo). La confiabilidad interobservador se estimó aplicando coeficiente de correlación intraclase. Finalmente, se evaluaron puntos de corte construyendo curvas ROC. El mejor punto de corte fue 24 puntos (área bajo la curva de 93,4 %). La validez de contenido fue de 80-100 % para todos los elementos incluidos. Validez de criterio y constructo fueron estadísticamente significativos ($p<0,05$). La confiabilidad interobservador fue de 0,96. Se diseñó y validó una escala para medir el CM de EPD.

PALABRAS CLAVE: Calidad metodológica; Diagnóstico; 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.

Bossuyt, P. M.; Reitsma, J. B.; Bruns, D. E.; Gatsonis, C. A.; Glasziou, P. P.; Irwig, L. M.; Lijmer, J. G.; Moher, D.; Rennie, D.; de Vet, H. C. & Standards for Reporting of Diagnostic Accuracy. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Standards for Reporting of Diagnostic Accuracy. *Clin. Chem., 49(1)*:1-6, 2003.

Cartes-Velásquez, R. A.; Manterola, C.; Aravena, P. & Moraga, J. Reliability and validity of MINCIR scale for methodological quality in dental therapy research. *Braz. Oral Res., 28*:1-5, 2014.

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

Cook, C.; Cleland, J.; Hegedus, E.; Wright, A. & Hancock, M. The creation of the diagnostic accuracy quality scale (DAQS). *J. Man. Manip. Ther., 22(2)*:90-6, 2014.

du Prel, J. B.; Röhrig, B. & Blettner, M. Critical appraisal of scientific articles. Part 1 of a series on evaluation of scientific publications. *Dtsch. Arztebl. Int., 106(7)*:100-5, 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.

George, D.; & Mallery, P. *SPSS for Windows step by step: A simple guide and reference*. 11.0 update. 4th ed. Boston, Allyn & Bacon. 2003.

Joyce, K. M.; Joyce, C. W.; Kelly, J. C.; Kelly, J. L. & Carroll, S. M. Levels of evidence in the plastic surgery literature: A citation analysis of the top 50 'classic' papers. *Arch. Plast. Surg., 42(4)*:411-8, 2015.

Leeflang, M.; Reitsma, J.; Scholten, R.; Rutjes, A.; Di Nisio, M.; Deeks, J. & Bossuyt, P. Impact of adjustment for quality on results of metaanalyses of diagnostic accuracy. *Clin. Chem., 53(2)*:164-72, 2007.

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.; 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.; 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.

Ochodo, E. A.; van Enst, W. A.; Naaktgeboren, C. A.; de Groot, J. A. H.; Hooft, L.; Moons, K. G. M.; Reitsma, J. B.; Bossuyt, P. M. & Leeflang, M. M. G. Incorporating quality assessments of primary studies in the conclusions of diagnostic accuracy reviews: a cross-sectional study. *B. M. C. Med. Res. Methodol., 14*:33, 2014.

Oliveira, M. R. F.; Gomes, A de C. & Toscano, C. M. QUADAS and STARD: evaluating the quality of diagnostic accuracy studies. *Rev. Saúde Pública, 45(2)*:416-22, 2001.

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

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

Wald, N. J. & Bestwick, J. P. Is the area under an ROC curve a valid measure of the performance of a screening or diagnostic test? *J. Med. Screen., 21(1)*:51-6, 2014.

Whiting, P. F.; Rutjes, A. W.; Westwood, M. E.; Mallett, S.; Deeks, J. J.; Reitsma, J. B.; Leeflang, M. M.; Sterne, J. A.; Bossuyt, P. M. & QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann. Intern. Med., 155(8)*:529-36, 2011.

Whiting, P. F.; Weswood, M. E.; Rutjes, A. W.; Reitsma, J. B.; Bossuyt, P. N. & Kleijnen, J. Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. *B. M. C. Med. Res. Methodol., 6*:9, 2006.

Whiting, P.; Rutjes, A. W.; Reitsma, J. B.; Bossuyt, P. M. & Kleijnen, J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *B. M. C. Med. Res. Methodol., 3*:25, 2003.

Wilson, F. R.; Pan, W. & Schumsky, D. A. Recalculation of the critical values for Lawshe's content validity ratio. *Meas. Eval. Couns. Dev., 45(3)*:197-210, 2012.

Corresponding author:
Prof. Dr. Carlos Manterola
Department of Surgery
Universidad de La Frontera
Avenida Francisco Salazar 01145
Temuco
CHILE

E-mail: carlos.manterola@ufrontera.cl

Received: 21-12-2017
Accepted: 16-03-2018