anales de psicología, 2012, vol. 28, nº 2 (mayo), 617-628
http://dx.doi.org/10.6018/analesps.28.2.148911

ISSN edición impresa: 0212-9728. ISSN edición web (http://revistas.um.es/analesps): 1695-2294

Methodological quality assessment tools of non-experimental studies: a systematic review

Alexander Jarde*, Josep-Maria Losilla, and Jaume Vives

Universitat Autònoma de Barcelona

Título: Instrumentos de evaluación de la calidad metodológica de estudios no experimentales: una revisión sistemática

Resumen: La evaluación de la calidad metodológica de los estudios primarios en una revisión sistemática es importante para garantizar la validez y fiabilidad de sus resultados, pero no existe acuerdo sobre qué instrumento debería usarse para hacerlo. Nuestro objetivo es analizar los instrumentos de medida utilizados en psicología y las ciencias de la salud para la valoración de estudios de cohortes, de casos y controles, y transversales. Se realizó una revisión sistemática usando 5 bases de datos y Google®. Para analizar el contenido de los instrumentos se definieron 6 dimensiones de calidad en base a guías de comunicación, bibliografía de referencia y estudios similares. Se identificaron y analizaron 74 instrumentos. Pocos indicaban su fiabilidad (20%) o validez (14%). Las dimensiones consideradas con más frecuencia fueron Obtención de datos (71.6%), Selección (67.6%), Análisis de datos y estadística (67.7%) y Medición (58.1%). Sólo un 35.1% consideraron Representatividad, y un 6.8% considera la Financiación. Pese a los puntos fuertes diseminados en los diferentes instrumentos, no hay ninguno que se pueda recomendar sin reservas. Un instrumento de medida para valorar la calidad metodológica de estudios no experimentales debería seguir un proceso de desarrollo estandarizado, pero previamente es necesario un acuerdo sobre qué dimensiones debería evaluar.

Palabras clave: estudios no experimentales; calidad metodológica; instrumentos de medida de la calidad; revisión sistemática.

Abstract: The evaluation of the methodological quality of primary studies in systematic reviews is of great importance in order to guarantee the validity and reliability of their results, but there is no agreement on which tool should be used. Our aim is to analyze the tools proposed so far for the assessment of cohort, case-control, and cross-sectional studies in psychology and health sciences. A systematic review was performed using 5 electronic databases and Google®. In order to analyze the tools' content, 6 domains of quality were defined based on reporting guidelines, the established bibliography, and previous similar studies. 74 tools were identified and analyzed. Few reported their reliability (20%) or validity (14%). The most frequently addressed content domains were Data collection (71.6%), Selection (67.6%), Statistics and data analysis (67.6%), and Measurement (58.1%); only 35.1% addressed Representativeness, and 6.8% addressed Funding. Despite the strengths we found scattered among the tools, there is no single obvious choice if we had to make any recommendation. Methodological quality assessment tools of non-experimental studies should meet standardized development criteria, but previously it is necessary to reach an agreement on which content domains they should take into account.

Key words: non-experimental studies; methodological quality; quality assessment tools; systematic review.

Introduction

Nowadays, the huge amount of information and the rate of publication make systematic reviews a crucial tool for researchers and health care providers (Martin, Pérez, Sacristán, & Álvarez, 2005; Wells & Littell, 2009). Although the inclusion of experiments in systematic reviews is well established, the inclusion of non-experimental studies is still under debate (Harden et al., 2004). However, much of clinical and public health knowledge is provided by non-experimental studies, and the area of psychology is not an exception. Indeed, about nine of ten research papers published in clinical journals are non-experimental studies, mainly cohort, case-control, and cross-sectional designs (Glasziou, Vandenbroucke, & Chalmers, 2004; Vandenbroucke et al., 2007) and a similar rate or even higher might be assumed in psychology. In fact, these designs are often the most efficient ones to answer certain questions and even may be the only practicable method of studying certain problems (Mann, 2003).

In a systematic review, the most difficult source of bias to control for is the low methodological quality of the selected studies. If the primary studies are flawed, then the conclusions of systematic reviews cannot be trusted. On the other hand, the quality scales for assessing primary studies greatly differ from one another and reach different conclusions about the quality of those studies (Jüni, Altman, & Egger, 2001; Jüni, Witschi, Bloch, & Egger, 1999; Valentine & Cooper, 2008). The Cochrane Collaboration suggests a tool for assessing susceptibility to bias which, according to its high frequency of use, could be considered as a standard for experiments (randomized controlled trials, RCT) in healthcare research (Higgins & Green, 2008). However, there is no consensus on which tool is the most appropriate to evaluate non-experimental studies (Sanderson, Tatt, & Higgins, 2007).

Considering the importance this type of studies have in clinical and public health knowledge in general, and in psychology in particular, and also considering the relevance of including them in systematic reviews in these areas, it becomes evident that there is a need of agreement about which tool to use to assess their methodological quality. More details about the current issues being debated around the use of quality scales to assess non-experimental studies can be found in Wells and Littell (2009).

Three systematic reviews focused on quality evaluation tools for non-experimental studies have been published up to date (Deeks et al., 2003; Sanderson et al., 2007; West et al., 2002). Both Deeks et al. (2003) and West et al. (2002) recommend six tools, but only coincide on half of them. In the most recent systematic review, Sanderson et al. (2007) conclude that there is no single obvious candidate tool for assessing quality of non-experimental studies. There is one important aspect that all previous systematic reviews agree on: most of the existing tools have not been developed using standard psychometric techniques. Although the concrete steps of these techniques differ in more or less degree

*** Dirección para correspondencia [Correspondence address]**: Alexander Jarde. Dpt. de Psicobiologia i de Metodologia de les CC. de la Salut; Facultat de Psicologia; Campus de la Universitat Autònoma de Barcelona; 08193. Cerdanyola del Vallès (Barcelona, Spain). E-mail: A.Jarde@gmail.com

among the reviews, they can be arranged with the following steps (Streiner & Norman, 1991): (a) The construct to be measured (in our case, "methodological quality") has to be operationally defined, (b) items have to be generated and/or selected, (c) some kind of pretesting of the items has to be done, and, once the tool is built, (d) its reliability and validity has to be assessed.

On the other hand, a remarkable aspect when comparing these reviews is the different interpretation of the concept of "methodological quality". We consider that a good approximation to this concept is that of "susceptibility to bias" as pointed out by the "STrengthering the Reporting of OBservational studies in Epidemiology" (STROBE) guidelines, developed by an international collaboration of epidemiologists, statisticians and journal editors and which is supported by many journals and organizations like the Annals of Behavioral Medicine, the World Health Organization Bulletin, and the Cochrane Collaboration (Vandenbroucke et al., 2007). Another relevant reporting guideline has recently been suggested by the American Psychology Association (APA) in its Publication Manual (APA, 2010): the Journal Article Reporting Standards (JARS), which addresses thoroughly the reporting of experimental and quasi-experimental studies but only partially those studies belonging to the non-experimental research.

The objective of our study is to carry out a systematic review of the tools proposed so far for the assessment of the methodological quality of studies with cohort, case-control, and cross-sectional designs in health sciences and, particularly, in psychology. The specific field of psychology has only been included in the review made by Deeks et al. (2003). Our revision takes into account the three mentioned research designs regardless of the topical focus of the study; in this sense, only the review of Sanderson et al. (2007) did not exclude any of these designs. For each tool, our revision extracts detailed information regarding the different stages of the tool's development; only the review of West et al. (2002) gives details of the whole tool's development process.

Method

Five electronic databases (Medline, Psycinfo, Cinahl, Cochrane Library and Dissertation Abstracts International) were searched for eligible studies published up to the beginning of the year 2010 (terms used in Medline can be found in Table 1). The search was not limited by language or by publication date. In an effort to capture those studies of interest not indexed by the chosen databases we also conducted an internet search using Google (http://www.google.com) with the results limited to the first 300 links.

Table 1. Keywords Used in Medline Corresponding to Each Search Element.

Search element	Keywords used in Medline
CREATION	develop*, elaborat*, "construct", "construction", "adapt", "adaptation", "proposal"
INSTRUMENT	checklist*, scale*, instrument*, tool*, "appraisal"
ASSESSMENT	assess*, evaluat*, measur*, "rate", "rating"
OBJECTIVE	"quality", "evidence", bias*, "confound", "confounding", "strength of", "validity"
STUDY	cohort stud*, follow-up stud*, case-control stud*, cross-sectional stud*, observational stud*, non-experimental stud*, epidemiologic stud*, "Cohort Studies"[Mesh], "Follow-Up Studies"[Mesh], "Case-Control Studies"[Mesh], "Cross-Sectional Studies"[Mesh], "Epidemiologic Studies"[Mesh].
APPLICATION	systematic*, review*, overview*, select*, search*, "look for", "find"

Note. Search elements were connected using the following structure: CREATION & INSTRUMENT & ASSESSMENT & OBJECTIVE & STUDY & APPLICATION. Keywords forming each search element were connected using "or". Keywords followed by [Mesh] are terms of the Medical Subject Heading.

Any published or unpublished document was eligible if it described a quality assessment tool applicable to cohort, case-control or cross-sectional studies. A tool was defined as any structured system of questions with a given set of possible answers. The tool could be itself the main aim of the publication or be included in the context of a systematic review. Tools based on other ones previously published were included as long as they added or modified the content of the original tool.

Search results were first filtered by title and abstract and, after that, the references of the remaining articles were reviewed. After this process there were 197 documents eligible. The full text of these documents was read by two of the authors (AJ and JML) independently and checked for the inclusion criteria. Differences of opinion were resolved by discussion or by the third author (JV). Finally, 74 results were included in this review. Figure 1 shows the whole search and the selection process in detail.

The included documents presented at least one tool each. When several tools were presented (e.g., different tools for cohort and for cross-sectional studies) each tool was considered independently. For each tool, two of the authors (AJ and JV, with differences of opinion resolved by discussion or by the third author, JML) extracted independently information about: (a) overall characteristics (number of items, designs addressed, type of tool and assessment), (b) information about the tool's development process (definition of the concept of "quality", item selection, previous pilot study, reliability analysis and validity analysis), and (c) which essential domains of methodological quality were assessed. A computerized data extraction form and its detailed guideline were developed to increase the reliability and the replicability of the whole process[1].

[1] The database and the extraction manual are available from the authors upon request.

Regarding the domains of methodological quality assessed by each tool, we consider that it is a key aspect that affects the validity and interpretability of our results, and therefore require a detailed justification. There is still very little empirical basis on which domains of quality affect to a major extent the validity of the results of the evaluated studies (West et al., 2002), so we defined six key domains based on three points. On one hand, we started from two widely supported reporting guidelines: the STROBE statement (Vandenbroucke et al., 2007), which is widely endorsed in health science, and the JARS (APA, 2010). On the other hand, we related our domains with the four inferential validity classes recognized throughout the social sciences (Shadish, Cook, & Campbell, 2002; Valentine & Cooper, 2008). Finally, we also crosschecked our domains against the ones proposed by the previous reviews, as they also have been developed by methodological experts.

It is important to differentiate the assessment of a study and the evaluation of an assessment tool. In this sense, our work is not intended to establish the necessary items for the assessment of individual studies, but to appraise whether the assessment tools proposed so far take into account the essential domains of methodological quality. So, with this in mind and on the basis of the three points aforementioned, we developed our six domains of quality.

1. Representativeness. Participants and non-participants are comparable on all important characteristics, including the sampled moments and situations, so that the selected sample properly represents the target study population. In order to identify a representative group of participants it is often necessary that, besides the participant's characteristics, the different moments and situations are taken into account during the sampling procedure (Shaughnessy, Zechmeister, & Zechmeister, 2009).

The information needed to assess this domain is present in several of the STROBE statement items, but especially in item 5 ("Setting"). The JARS requests this information in its "Sampling procedures" section. Of the previous reviews, the ones by Deeks et al. (2003) and by Sanderson et al. (2007) also somehow deal with this domain. We considered that a tool dealt with this domain (which does not necessarily involve that it is totally covered) if it had items to appraise the justification of the sample representativeness or regarding the similarity between participants and non-participants.

2. Selection. The different groups of participants are comparable on all important characteristics except on the variables under study. In general, groups under comparison should have a similar distribution of characteristics (being or not under direct investigation). If groups differ from each other in a systematic way, the interpretation of results may become confused (Avis, 1994; Fowkes & Fulton, 1991; Higgins & Green, 2008). A variable not directly under study becomes a confounding factor if it is associated with the outcome under study and if its distribution is different between the groups compared. They can be understood as a problem of comparability with its origin linked to the impossibility of making a random assignment of participants (Hernández-Avila, Garrido, & Salazar-Martínez, 2000; Mann, 2003; Shaughnessy et al., 2009). Efforts can be done to control confounding by design using control techniques as matching or restriction in order to balance the groups under comparison (Shadish et al., 2002).

The STROBE statement requests the necessary information to assess this domain in item 6 ("Participants") and the JARS in its "Participant characteristics" section. All the previous reviews deal with this domain with more or less items. We considered that a tool dealt with this domain (which does not necessarily involve that it is totally covered) if it had items to appraise if eligibility criteria are clearly defined, as well as balancing criteria, and if they are applied in the same way to all groups.

3. Measurement. The instruments used to collect the data are appropriate (valid and reliable). The choice of one instrument or another to measure the variables under study should be based not only on its reliability and validity, but also on the definition of the construct it measures (Carretero-Dios & Pérez, 2007).

Item 8 ("Data sources/measurement") of the STROBE statement and the "Measures and covariates" section of the JARS demand the necessary information to assess this domain, which is taken into account in all previous reviews. To consider that a tool addressed this domain (which does not necessarily involve that it is totally covered) it should contain items that forced to judge the appropriateness of the measurement tools.

4. Data Collection. The comparability of the groups and the data quality are not affected by threats that may appear during the data collection and management. Other threats to validity may appear if there are systematic differences between groups in how the information is collected (Hernández-Avila et al., 2000; Sica, 2006). Knowing the purpose and objectives of the study is a common source of bias during data collection, so masking of study participants and researchers is important. Total masking is not feasible in many studies, but it is necessary to consider how this might put the results in doubt (Fowkes & Fulton, 1991; Kopec & Esdaile, 1990; Shadish et al., 2002).

The information needed to assess this domain is present in the item 9 of the STROBE statement ("Bias"). There is a section ("Masking") in the JARS that would be related to this domain although it only appears in its design-specific modules of randomized experiments and quasi-experiments. All previous reviews somehow assess this domain. We considered that a tool dealt with this domain (which does not necessarily involve that it is totally covered) if it had items checking if some kind of masking was done to the participants and/or the researchers involved. Items checking for other methods, different from masking, to control these threats to comparability and data quality (e.g., interviewer bias or memory bias) also made us consider that the tool dealt with this domain.

5. Statistics and Data Analysis. The different groups remain comparable despite incomplete data (due to missing data or loss to follow-up) and potentially confounding variables are controlled for in statistical analysis. Confounding factors may also be minimized by some form of stratification or adjustment procedure in the analysis. This is especially relevant if the confounding variables were not controlled for by design. The potentially confounding variables must have been measured, though, so it is necessary that researchers think carefully about them beforehand (Fowkes & Fulton, 1991; Vandenbroucke et al., 2007).

In cohort studies there are many reasons why subjects cannot be followed up completely; although this does not necessarily lead to bias, careful analysis is required to rule it out; and this not only applies to the proportion of drop outs but also to the reason why (Avis, 1994; Fowkes & Fulton, 1991; Sica, 2006; Vandenbroucke et al., 2007).

It is of special relevance to take missing data into account, since they can reduce the legitimacy of the results and, if participants with missing data are not representative of the whole sample, bias may arise (Fowkes & Fulton, 1991; Vandenbroucke et al, 2007).

Item 12 ("Statistical methods") of the STROBE statement demands the necessary information to assess this domain, but there is no related section in the JARS. The previous reviews by Sanderson et al. (2007) and by West et al. (2002) take this domain into account. A tool was considered to deal with this domain (which does not necessarily involve that it is totally covered) if it had at least one item checking whether potentially confounding variables were controlled for in the statistical analysis, groups were comparable regarding the number and characteristics of subjects with incomplete data, or incomplete data affected the compared groups in the same way.

6. Funding. The sources of funding and possible conflicts of interests have not influenced the study. Several studies show strong associations between the source of funding and the conclusions of research articles. Funding may affect the study design, choice of exposures, outcomes, statistical methods, and selection of outcomes and studies for publication (Vandenbroucke et al., 2007).

The STROBE statement requests to publish information regarding funding in its item 22 ("Funding") while the JARS does so in its "Title and title page" section. All previous reviews except the one by Deeks et al. (2003) address this domain. We considered that a tool dealt with this domain if it included any item checking for the study funding or conflicts of interests.

Results

The 74 analyzed tools have five to 85 items, with a median of 15 (interquartile range = 15). Table 2 shows the main characteristics of the analyzed tools.

Table 2. Main Characteristics of the Analyzed Tools.

				Tool development						Content's domains					
Author, Year	Design	Type	Items	Qual.	Adapt.	Emp.	Pilot	Rel.	Val.	Rep.	Sel.	Meas.	D.col.	Stat.	Fund.
Angelillo and Villari, 1999	Coh, CC, CS	Chl	24		x					x	x	x	x	x	
Ariëns, Van Mechelen, Bongers, Bouter, and Van der Wal, 2000	Coh, CC, CS	Chl	22					x		x		x	x	x	
Atluri, Datta, Falco, and Lee, 2008	Coh, CC, CS	Chl	26		x					x		x	x	x	x
Avis, 1994	Coh, CC, CS	Chl	24		x					x	x		x	x	
Berra, Elorza-Ricart, Estrada, and Sanchez, 2008	CS	Chl	27		x					x	x	x	x	x	x
Bhutta, Cleves, Casey, Cradock, and Anand, 2002	CC	Scl	6					x		x					
Bishop et al., 2009	CS	Chl	17		x			x		x		x	x	x	
Blagojevic, Jinks, Jeffery, and Jordan, 2010	Coh, CC	Chl	15		x			x		x	x	x		x	
Borghouts, Koes, and Bouter, 1998	Coh, CC	Chl	13	x	x			x							
Buckingham et al., 2003a	Coh	Chl	9							x	x	x	x	x	
Buckingham et al., 2003b	CC	Chl	9							x			x	x	
Cameron et al., 2000	Coh	Chl	9			x				x	x		x	x	
Campbell and Rudan, 2002	CC	Chl	13							x		x		x	
Campos-Outcalt, Senf, Watkins, and Bastacky, 1995	Coh, CC, CS	Scl	9											x	
Carruthers, Larochelle, Haynes, Petrasovits, and Schiffrin, 1993	Coh	Chl	6											x	

				Tool development						Content's domains					
Author, Year	Design	Type	Items	Qual.	Adapt.	Emp.	Pilot	Rel.	Val.	Rep.	Sel.	Meas.	D.col.	Stat.	Fund.
Critical Appraisal Skills Programme Español (CASPe), 2008a	CC	Chl	11		x					x		x	x	x	
Critical Appraisal Skills Programme Español (CASPe), 2008b	Coh	Chl	11							x	x		x	x	
Centre for Evidence-Based Medicine, 2004	Coh	Chl	7							x			x	x	
Cho and Bero, 1994	Coh, CC	Chl	23	x	x		x	x	x	x	x		x	x	
Cole and Hudak, 1996	Coh	Chl	6										x	x	
Corrao, Bagnardi, Zambon, and Arico, 1999	Coh, CC	Chl	15								x	x	x		
Cowley, 1995	Coh	Chl	13							x		x	x	x	x
Downs and Black, 1998	Coh, CC, CS	Chl	27		x		x	x	x	x	x	x	x	x	
DuRant, 1994	CC, CS	Chl	62							x	x	x	x	x	
Effective Practice, Informatics and Quality Improvement (EPIQ), 2008a	Coh	Chl	24							x	x	x	x	x	
Effective Practice, Informatics and Quality Improvement (EPIQ), 2008b	CC	Chl	22							x	x	x	x	x	
Effective Public Health Practice Project (EPHPP), 2009	Coh, CC	Chl	21							x		x	x	x	
Esdaile and Horwitz, 1986	Coh, CC	Chl	6							x					
Federal Focus, 1996	Coh, CC	Chl	33							x		x	x	x	
Fowkes and Fulton, 1991	Coh, CC, CS	Chl	22	x						x	x	x	x	x	
Gardner, Machin, and Campbell, 1986	Coh, CC, CS	Chl	12												
Genaidy et al., 2007	Coh, CC, CS	Chl	43		x	x	x	x	x	x		x	x	x	
Glasgow University, 2009	Coh, CC	Chl	10		x					x		x	x		
Greer, Mosser, Logan, and Halaas, 2000	Coh, CC	Chl	10		x					x					
Gyorkos et al., 1994	Coh, CC, CS	Chl	6										x		
Hadorn, Baker, Hodges, and Hicks, 1996	Coh	Chl	32	x	x						x		x	x	
Khan, ter Riet, Glanville, Sowden, and Kleijnen, 2001	Coh, CC	Chl	25	x						x	x	x	x	x	
Kreulen, Creugers, and Meijering, 1998	Coh, CC, CS	Chl	15		x			x					x		
Krogh, 1985	Coh, CC, CS	Chl	11							x					
Kwakkel, Wagenaar, Kollen, and Lankhorst, 1996	Coh	Chl	11		x							x		x	
Laupacis, Wells, Richardson, and Tugwell, 1994	Coh	Chl	7								x			x	
Levine et al., 1994	Coh, CC	Chl	7							x					
Lichtenstein, Mulrow, and Elwood, 1987	CC	Chl	20			x	x			x			x		
Liddle, Williamson, and Irwig, 1996	Coh, CC	Chl	10		x		x					x	x	x	
Littenberg et al., 1998	Coh, CC, CS	Chl	5										x		
Loney and Stratford, 1999	CS	Chl	9								x	x	x		
López de Argumedo et al., 2006	Coh	Chl	60				x		x	x		x	x	x	x

				Tool development						Content's domains					
Author, Year	Design	Type	Items	Qual.	Adapt.	Emp.	Pilot	Rel.	Val.	Rep.	Sel.	Meas.	D.col.	Stat.	Fund.
Margetts et al., 1995 (CC)	CC	Mix	24					x	x	x				x	
Margetts et al., 1995 (Coh)	Coh,	Chl	19					x	x					x	
Margetts, Vorster, and Venter, 2002	Coh, CC, CS	Chl	22									x			
New Zealand Guidelines Group, 2001	Coh, CS	Chl	25							x	x	x	x		
Nguyen, Bezemer, Habets, and Prahl-Andersen, 1999	Coh, CC, CS	Scl	18									x	x	x	
Parker, 2006	Coh, CC	Chl	29							x		x	x		
Pérez-Rios et al., 2009	Coh, CC	Chl	5		x			x				x		x	
Rangel, Kelsey, Colby, Anderson, and Moss, 2003	CC	Chl	23		x		x	x				x	x		
Reed et al., 2007	Coh, CC, CS	Chl	10		x	x	x	x	x						
Reisch, Tyson, and Mize, 1989	Coh, CC	Chl	85					x		x		x	x	x	x
Scottish Intercollegiate Guidelines Network, 2008 (CC)	CC	Chl	13		x					x		x	x	x	
Scottish Intercollegiate Guidelines Network, 2008 (Coh)	Coh	Chl	16		x					x		x	x	x	
Solomon, Bates, Panush, and Katz, 1997	Coh	Chl	11							x		x	x	x	
Spitzer et al., 1990	Coh, CC, CS	Chl	32		x					x	x	x	x	x	
Steinberg et al., 2000	Coh, CC, CS	Chl	24							x			x	x	
Stock, 1991	Coh, CC, CS	Chl	7							x	x	x	x	x	
The Joanna Briggs Institute, 2008	Coh, CC	Chl	9								x	x	x		
Tseng, Breau, Fesperman, Vieweg, and Dahm, 2008	Coh	Chl	45		x		x					x	x		
van der Windt et al., 2000	Coh, CC, CS	Chl	25		x					x		x	x	x	
Vitali and Randolph, 2005	Coh, CC	Chl	12							x			x	x	
Weightman, Mann, Sander, and Turley, 2004	Coh, CC, CS	Chl	25		x					x			x	x	
Wells et al., 2009 (CC)	CC	Chl	8					x	x	x	x		x		
Wells et al., 2009 (Coh)	Coh	Chl	8					x	x	x	x			x	
Welsh Child Protection Systematic Review Group, 2006	Coh, CC, CS	Chl	37		x					x			x		
Wong, Cheung, and Hart, 2008	CC, CS	Chl	5		x		x	x		x	x	x		x	
Zaza et al., 2000	Coh, CC, CS	Chl	20	x	x		x		x	x		x	x	x	
Zola et al., 1989	Coh, CC, CS	Chl	13		x										

Note. Design = Design to which the tool is applicable; Coh = Cohort design; CC = Case-Control design; CS = Cross-sectional design; Chl = Checklist (items with categorical answers); Scl = Scale (items with numeric answers); Mix = Items with categorical and numeric answers; Items = Number of items; Qual. = Definition of the concept of "quality"; Adapt. = Items adapted from other tools; Emp. = Empirical development of the items; Pilot = Pilot study; Rel. = Reliability analysis; Val. = Validity analysis; Rep. = Representativeness; Sel. = Selection; Meas. = Measurement; D.col. = Data collection; Stat. = Statistics and data analysis; Fund. = Funding.

Of the analyzed tools, 28 (37.8%) are specific for one type of design, while the rest of them can be applied to two or more of the considered designs. While most of all tools are applicable to cohort studies (61 tools, 82.4%) and case-control studies (53 tools, 71.6%), much less are applicable to cross-sectional studies (30, 40.5%). Details on the applicability of the tools can be found in Table 3.

We found that 70 (94.6%) tools were checklists (a simple list of items). Forty-three (58.1%) tools apply some kind of summary score and eight of them (10.8%) use a subjective categorical evaluation. Details on the descriptive characteristics of the tools can be found in Table 4.

Table 3. Number (%) of tools applicable to each design

Research designs			
Cohort	Case-Control	Cross-sectional	*n* (%)
x	x	x	24 (32.4%)
x	x		19(25.7%)
x		x	1 (1.4%)
	x	x	2 (2.7%)
x			17 (23.0%)
	x		8 (10.8%)
		x	3 (4.1%)
61 (82.4%)	53 (71.6%)	30 (40.5%)	

Table 4. Descriptive information of the tools

Tool's description	*n*	%
Type of tool		
Checklist	70	94.6%
Scale	3	4.1%
Mixed	1	1.4%
Summary score		
None	31	41.9%
Direct calculation	28	37.8%
Weighted calculation	7	9.5%
Categorical	8	10.8%

Note. A tool was considered a checklist if its items had categorical answers, a scale if its items were answered numerically, and of mixed type if it had items with both categorical and numeric answers.

In general, the development of the analyzed tools does not meet the standardized development criteria for a measuring tool (Carretero-Dios & Pérez, 2007; Streiner & Norman, 1991). In fact, only five (6.8%) tools discuss the concept of "quality". Less than half of the tools (32, 43.2%) inform about the origin of the items, being most of them adapted from other tools (29, 39.2%). Only in four cases (5.4%) the items were developed using an empirical approach (e.g., Delphi technique). It is worth mentioning that 85.1% of the tools (63) do not test any pilot version, 75.7% of the tools (56) do not make any kind of reliability analysis, and 86.5% (64) of them do not assess its validity.

Taking a closer look at how each of the 74 tools was developed, only five cover at least four of the five aspects we relate to a proper development of a tool (Cho & Bero, 1994; Downs & Black, 1998; Genaidy et al., 2007; Reed et al., 2007; Zaza et al., 2000). The one presented by Cho and Bero (1994) is the only tool that covers all five aspects evaluated (Table 5).

In respect to their contents, twenty-six tools (35.1%) assess the representativeness of the sample and 50 (67.6%) deal with the selection of participants and the comparability of the groups. Forty-three tools (58.1%) require assessing the measurement of the variables and 53 tools (71.6%) take the threats to validity during data collection into account. Finally, 50 tools (67.6%) assess the control for confounding or consider missing data or loss to follow-up in the statistics and data analysis, and five (6.8%) check for bias due to funding. Table 6 shows in detail the characteristics of the tool's content.

Table 5. Tools covering at least four of the desirable aspects during its development.

Tool	Discussion of the concept "quality"	Item selection	Pilot study	Reliability tests	Validity tests
Cho and Bero, 1994	x	x	x	x	x
Downs and Black, 1998		x	x	x	x
Genaidy et al., 2007		x	x	x	x
Reed et al., 2007		x	x	x	x
Zaza et al., 2000	x	x	x		x
Total	5 (7%)	29 (39%)	11 (15%)	18 (24%)	10 (14%)

Table 6. Number of tools assessing each content domain for each study design.

Content domain	Cohort (% with n=60)	Case-Control (% with n=52)	Cross-sectional (% with n=30)	Overall (% with n=74)
Representativeness	19 (31.7%)	15 (28.8%)	12 (40%)	26 (35.1%)
Selection	39 (65.5%)	37 (71.2%)	19 (63.3%)	50 (67.6%)
Measurement	33 (55.0%)	30 (57.7%)	18 (60%)	43 (58.1%)
Data collection	43 (71.7%)	36 (69.2%)	23 (76.7%)	53 (71.6%)
Statistics and data analysis	41 (68.3%)	31 (59.6%)	19 (63.3%)	50 (67.6%)
Funding	4 (6.7%)	2 (3.8%)	2 (6.7%)	5 (6.8%)

As shown in Table 7, there are 11 tools (14.9%) that somehow assess the six content related domains that we consider essential or all except for the domain Funding. We applied a more demanding filter to these tools, considering separately both the incomplete data (loss to follow-up and missing data) and confusion management in the statistical analysis. Only five tools pass this new filter (Berra, Elorza-Ricart, Estrada, & Sanchez, 2008; Buckingham, Fisher, & Saunders, 2003a; Downs & Black, 1998; DuRant, 1994; Khan, ter Riet, Glanville, Sowden, & Kleijnen, 2001).

The first one, presented by Berra et al. (2008), is applicable to cross-sectional studies only and assesses all six domains. It was developed based on previous tools and on the STROBE statement, although no reliability or validity data is given. The tool is structured in eight topics containing one to five items each, with 27 in total. Each item has to be marked in how far the considered aspect has been achieved (*Very well*, *Well*, *Regular*, *Bad*), or if information is missing or if the item is not applicable. Furthermore, the tool demands

its user to make an evaluation of each topic and of the whole study.

The worksheet for using an article about prognosis of the Evidence Based Medicine Toolkit (EBM Toolkit) of Buckingham et al. (2003a) assesses all six domains except Funding. As it is designed to support critical appraisal of studies, it is divided in three parts: The first one assesses the study validity, the second one the study results and the third one deals with the applicability of the results to the reader's patients. The study validity part has only five items, which have to be answered with *Yes*, *No* or *Can't tell*; but some of them include several questions (to be answered with *yes* or *no*) that expand the assessment of the item. It is adapted from a series of guideline articles for medical literature, but we could not find any more information about its development.

The tool proposed by Downs & Black (1998) also assesses all six domains except Funding. It is a checklist with 27 items applicable to experimental and non-experimental studies. Although it is presented as a methodological quality assessment tool, 10 items assess the study reporting and one item assesses the study's statistical power. The tool was developed based on bibliographic reviews and existing tools to assess experimental studies, and a pilot study was done previously. Data is given for its internal consistency, criterion validity, and test-retest and inter-rater reliability.

The tool proposed by DuRant (1994) has 62 items distributed in six topics and is applicable to case-control studies and cross-sectional studies (although it has other items to also assess experimental and quasi-experimental studies). It assesses all six domains except the one of Funding and there is no information about how it was developed.

Finally, Khan et al. (2001) give "some quality criteria for assessment of observational studies" for cohort studies and case-control studies (and case series) without presenting it as an assessment tool, so it is no surprise that no data is given about its development. It assesses all six domains except Funding using only 10 items for each design type.

Table 7. Domains covered by the highlighted tools.

		Content's domains						
Author, Year	Design	Rep.	Sel.	Meas.	Data col.	Statistics		Fund.
						Conf.	Inc. data	
Berra et al., 2008	CS	x	x	x	x	x	x	x
Buckingham et al., 2003a	Coh	x	x	x	x	x	x	
Khan et al., 2001	Coh, CC	x	x	x	x	x	x	
Downs & Black, 1998	Coh, CC, CS	x	x	x	x	x	x	
DuRant, 1994	CC, CS	x	x	x	x	x	x	
EPIQ[a], 2008a	Coh	x	x	x	x	x		
EPIQ[a], 2008b	CC	x	x	x	x	x		
Angelillo & Villari, 1999	Coh, CC, CS	x	x	x	x		x	
Fowkes & Fulton, 1991	Coh, CC, CS	x	x	x	x	x		
Stock, 1991	Coh, CC, CS	x	x	x	x	x		
Spitzer et al., 1990	Coh, CC, CS	x	x	x	x	x		

Note. Tools ordered by number of domains addressed and publication date. Rep. = Representativeness; Sel. = Selection; Meas. = Measurement; Data col. = Data collection; Conf. = Confusion controlled for in the statistical analysis; Inc. data = Incomplete data (lost to follow-up and missing data) considered in the statistical analysis; Fund = Funding; Coh = Cohort design; CC = Case-Control design; CS = Cross-sectional design.
[a] Effective Practice, Informatics and Quality Improvement

Discussion

As happened to Deeks et al. (2003) and West et al. (2002), our search in the different databases is the less productive information source, since most of the analyzed tools were found reviewing the references of the database search results (after filtering by title and abstract. See Figure 1). This can be explained by the fact that lots of tools are developed for specific systematic reviews, which makes its identification using a database search difficult (Sanderson et al., 2007). We are also aware that the keywords used to perform the Boolean search might have been too narrow, but we had to balance between strategies that were less likely to miss any relevant papers, yet retrieving a *manageable* number of results. Consequently, our search has probably not been exhaustive. We consider it sensible enough though, since we have located all the tools considered most relevant by previous similar systematic reviews.

The first remarkable conclusion of our systematic review is the ascertainment that most of the existing tools up to date have not been developed rigorously. In this sense, only one tool (Cho & Bero, 1994) covers the five criteria we consider important. This is worrying, since most of the analyzed tools are intended to be used in systematic reviews where rigorousness during the methodological quality assessment of the studies is a key point.

Focusing exclusively on their contents, our second conclusion is that there is no single obvious choice among the most comprehensive tools we have reviewed. In this sense, we agree with the results of the systematic review by Sander-

son et al. (2007). As aforementioned in the results section, only one tool takes into consideration all six content domains evaluated (Berra et al., 2008) and 10 more somehow appraise all except Funding. Of these 11 tools, only five pass our more demanding filter (using a more strict consideration of the Statistics and data analysis domain) (Berra et al., 2008; Buckingham et al., 2003a; Downs & Black, 1998; DuRant, 1994; Khan et al., 2001). Considering both the tool development and content domains assessed, only the tool proposed by Downs and Black (1998) reach the minimum requirements, but has some limitations in these respects, as for example the lack of a definition of the concept of quality or any item assessing the source of funding and conflicts of interests.

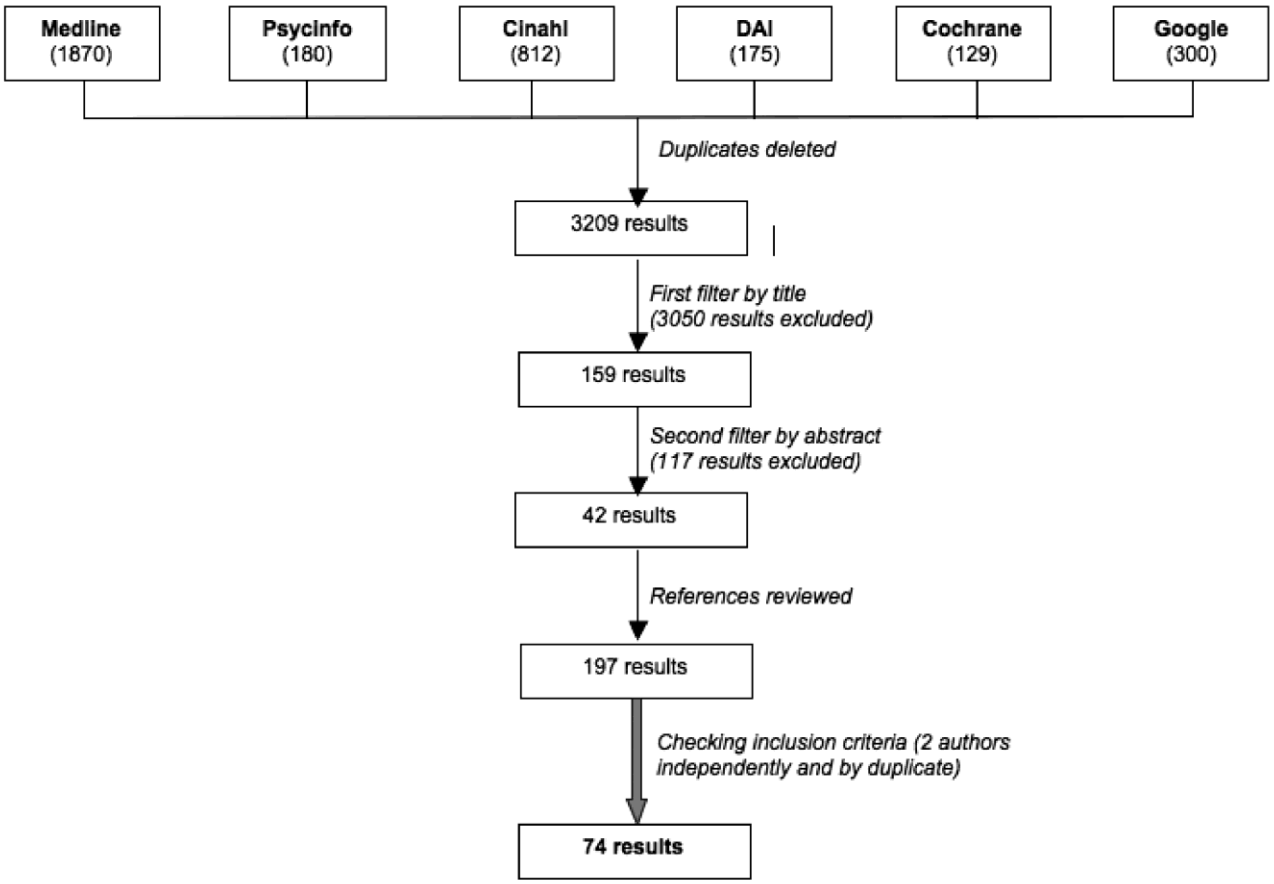

Figure 1. Search and selection process of all documents included in this review.

Finally, our third conclusion is that there is no agreement on which content domains should be taken into account in order to assess the methodological quality of non-experimental studies, which is reflected by the heterogeneity of the domains addressed by the reviewed tools. This is a key issue, and consequently an important previous step that has to be achieved. We consider that future studies should focus on it. In our review we have seen that the domains more frequently addressed are Selection, Statistics and data analysis, Data collection, and, in a lesser degree, Measurement. On the other hand, very few tools cover Funding, which is consistent with all previous reviews that take funding into account (Sanderson et al., 2007; West et al., 2002), and Representativeness (only addressed by one third of the analyzed tools), which is probably due to the fact that most authors do not include this aspect in their concept of methodological quality. These conclusions are applicable to all three study designs reviewed.

When trying to compare our results with the ones of previous systematic reviews, it becomes clear that it is not possible to do so with the data extracted in the review by Sanderson et al. (2007). Leaving aside the fact that they do not make any selection of acceptable or best tools, the procedure they follow consists of counting for each tool the number of items that are somehow related with any of their six domains. But we consider that when a tool has a high number of items related to a domain does not necessarily imply that the construct represented by that domain is correctly assessed. In contrast, Deeks et al. (2003) and West et al. (2002) qualitatively evaluate if the domains and concrete elements they consider essential are assessed. This is the procedure that we have followed, which makes the compari-

son of our results with the ones of Deeks et al. (2003) and West et al. (2002) feasible. So, when comparing their list of highlighted tools with ours, we find that there are only two tools that are recommended by Deeks et al. (2003) and West et al (2002), and that also assess all our six domains or all except Funding (Downs & Black, 1998; Spitzer et al., 1990). The main reason why the other tools recommended by either Deeks et al. (2003) or West et al. (2002) (or by both of them) are not highlighted in our review is because they do not address our domain Representativeness (Cowley, 1995; Effective Public Health Practice Project [EPHPP], 2009; Reisch, Tyson, & Mize, 1989; Scottish Intercollegiate Guidelines Network, 2008 [cohort designs and case-control designs]; Zaza et al., 2000). In two other cases where the domain Representativeness was addressed, other domains were missed (Wells et al., 2009 [cohort designs and case-control designs]). This is so considering that in some cases we have analyzed a more recent version of the tools than those evaluated by Deeks et al. (2003) and West et al. (2002).

Six of the remaining nine tools we have highlighted were developed after the previous reviews were published (Berra et al., 2008; Buckingham et al., 2003a; Effective Practice, Informatics and Quality Improvement [EPIQ], 2008a, 2008b; Khan et al., 2001). In addition, three tools were published prior to the date range used by West et al. (2002) in its search strategy (from year 1995 to 2000); two of them (DuRant, 1994; Fowkes & Fulton, 1991) are considered among the best tools by Deeks et al. (2003), while the other one (Stock, 1991) does not satisfy their criteria. Finally, the remaining tool that we have highlighted (Angelillo & Villari, 1999) was not retrieved in Deeks et al.'s (2003) review for some reason; it appears in West et al.'s (2002) review, where, although it is very well considered, it is not selected as one of the recommended tools.

We hope this review may be a step further in the path to the development as well as to the consensus of a quality assessment tool that may be applied in future systematic reviews using cohort, case-control and cross-sectional studies as its primary articles.

Acknowledgements.- This research was supported by Grant PSI2010-16270 from the Spanish Ministry of Science and Innovation (Spain).

References

American Psychological Association. (2010). Publication Manual of the American Psychological Association, Sixth Edition (6th ed.). Washington, DC: Author.

Angelillo, I. F., & Villari, P. (1999). Residential exposure to electromagnetic fields and childhood leukaemia: A meta-analysis. *Bulletin of the World Health Organization, 77*(11), 906-915.

Ariëns, G. A. M., Van Mechelen, W., Bongers, P. M., Bouter, L. M., & Van der Wal, G. (2000). Physical risk factors for neck pain. *Scandinavian Journal of Work, Environment & Health,* 26(1), 7-19.

Atluri, S., Datta, S., Falco, F. J., & Lee, M. (2008). Systematic review of diagnostic utility and therapeutic effectiveness of thoracic facet joint interventions. *Pain Physician, 11*(*5*), 611-629.

Avis, M. (1994). Reading research critically. II. An introduction to appraisal: Assessing the evidence. *Journal of Clinical Nursing, 3*(*5*), 271-277.

Berra, S., Elorza-Ricart, J. M., Estrada, M. D., & Sánchez, E. (2008). A tool for the critical appraisal of epidemiological cross-sectional studies. [Instrumento para la lectura crítica y la evaluación de estudios epidemiológicos transversales] *Gaceta sanitaria / S.E.S.P.A.S, 22*(5), 492-497.

Bhutta, A. T., Cleves, M. A., Casey, P. H., Cradock, M. M., & Anand, K. J. S. (2002). Cognitive and behavioral outcomes of school-aged children who were born preterm: A meta-analysis. *JAMA: The Journal of the American Medical Association, 288*(6), 728-737.

Bishop, F. L., Prescott, P., Chan, Y. K., Saville, J., von Elm, E., & Lewith, G. T. (2010). Prevalence of Complementary Medicine Use in Pediatric Cancer: A Systematic Review. *Pediatrics*, *125*(4), 768-776.

Blagojevic, M., Jinks, C., Jeffery, A., & Jordan, K. (2010). Risk factors for onset of osteoarthritis of the knee in older adults: a systematic review and meta-analysis. *Osteoarthritis and Cartilage*, *18*(1), 24-33.

Borghouts, J. A., Koes, B. W., & Bouter, L. M. (1998). The clinical course and prognostic factors of non-specific neck pain: A systematic review. *Pain, 77(*1), 1-13.

Buckingham, J., Fisher, B., & Saunders, D. (2003a). Worksheet for using an article about prognosis. *Evidence Based Medicine Toolkit*. Retrieved May 5, 2009, from http://www.ebm.med.ualberta.ca/

Buckingham, J., Fisher, B., & Saunders, D. (2003b). Worksheet for Using an Article About Causation of Harm. *Evidence Based Medicine Toolkit*. Retrieved May 5, 2009, from http://www.ebm.med.ualberta.ca/

Cameron, I., Crotty, M., Currie, C., Finnegan, T., Gillespie, L., Gillespie, W., et al. (2000). Geriatric rehabilitation following fractures in older people: A systematic review. *Health Technology Assessment, 4*(2).

Campbell, H., & Rudan, I. (2002). Interpretation of genetic association studies in complex disease. *The Pharmacogenomics Journal, 2*, 349-360.

Campos-Outcalt, D., Senf, J., Watkins, A. J., & Bastacky, S. (1995). The effects of medical school curricula, faculty role models, and biomedical research support on choice of generalist physician careers: A review and quality assessment of the literature. *Academic Medicine : Journal of the Association of American Medical Colleges, 70(7*), 611-619.

Carretero-Dios, H., & Pérez, C. (2007). Standards for the development and review of instrumental studies: Considerations about test selection in psychological research. *International Journal of Clinical and Health Psychology, 7*(3), 863-882.

Carruthers, S. G., Larochelle, P., Haynes, R. B., Petrasovits, A., & Schiffrin, E. L. (1993). Report of the canadian hypertension society consensus conference: 1. introduction. *Canadian Medical Association Journal, 149(*3), 289-293.

Centre for Evidence-Based Medicine (CEBM). *Critical appraisal worksheets.* Unpublished document. Retrieved April 4, 2009, from http://www.cebm.net/index.aspx?o=1913.

Cho, M. K., & Bero, L. A. (1994). Instruments for assessing the quality of drug studies published in the medical literature. *The Journal of the American Medical Association, 272*(2), 101-104.

Cole, D. C., & Hudak, P. L. (1996). Prognosis of nonspecific work-related musculoskeletal disorders of the neck and upper extremity. *American Journal of Industrial Medicine, 292,* 657-668.

Corrao, G., Bagnardi, V., Zambon, A., & Arico, S. (1999). Exploring the dose-response relationship between alcohol consumption and the risk of several alcohol-related conditions: A meta-analysis. *Addiction, 94(*10), 1551-1573.

Cowley, D. E. (1995). Prostheses for primary total hip replacement: A critical appraisal of the literature. *International Journal of Technology Assessment in Health Care, 11*(4), 770-778.

Critical Appraisal Skills Programme Español (CASPe). (2008a). Critical appraisal tools [herramientas para lectura crítica] - Case-Control studies. *CASPe.* Retrieved May 5, 2009, from http://www.redcaspe.org/herramientas/index.htm

Critical Appraisal Skills Programme Español (CASPe). (2008b). Critical appraisal tools [herramientas para lectura crítica] - Cohort studies. *CASPe*. Retrieved May 5, 2009, from http://www.redcaspe.org/herramientas/index.htm

Deeks, J. J., Dinnes, J., D'Amico, R., Sowden, A. J., Sakarovitch, C., Song, F., Petticrew, M., & Altman, D.G. (2003). Evaluating non-randomized intervention studies. *Health Technology Assessment, 7*(27), 1-173.

Downs, S. H., & Black, N. (1998). The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *Journal of Epidemiology and Community Health, 52*(6), 377-384.

DuRant, R. H. (1994). Checklist for the evaluation of research articles. *Journal of Adolescent Health, 15*(1), 4-8.

Effective Practice, Informatics and Quality Improvement (EPIQ). (2008a). *Critically appraised topics (CATs) checklists for quantative studies: Prognostic and Risk Factor Studies*. Retrieved January 9, 2010 from http://www.epiq.co.nz

Effective Practice, Informatics and Quality Improvement (EPIQ). (2008b). *Critically appraised topics (CATs) checklists for quantative studies: Case-control studies*. Retrieved January 9, 2010 from http://www.epiq.co.nz

Effective Public Health Practice Project (EPHPP). (2009). *Quality assessment tool for quantitative studies*. Retrieved January 9, 2010, from http://www.ephpp.ca/PDF/QATool.pdf

Esdaile, J. M., & Horwitz, R. I. (1986). Observational studies of cause-effect relationships: An analysis of methodologic problems as illustrated by the conflicting data for the role of oral contraceptives in the etiology of rheumatoid arthritis. *Journal of Chronic Diseases, 39*(10), 841-852.

Federal Focus (1996). Principles for evaluating epidemiologic data in regulatory risk assessment. Unpublished document. Retrieved January 9, 2010, from http://www.fedfocus.org/science/london-principles.html.

Fowkes, F. G., & Fulton, P. M. (1991). Critical appraisal of published research: Introductory guidelines. *British Medical Journal, 302*, 1136-1140.

Gardner, M. J., Machin, D., & Campbell, M. J. (1986). Use of check lists in assessing the statistical content of medical studies. *British Medical Journal, 292*, 810-812.

Genaidy, A. M., Lemasters, G. K., Lockey, J., Succop, P., Deddens, J., Sobeih, T., & Dunning, K. (2007). An epidemiological appraisal instrument - a tool for evaluation of epidemiological studies. *Ergonomics, 50*(6), 920-960.

Glasgow University. *Critical appraisal checklist for an article on harm or causation*. Unpublished document. Retrieved January 9, 2010, from http://www.gla.ac.uk/media/media_64043_en.pdf

Glasziou, P., Vandenbroucke, J. P., & Chalmers, I. (2004). Assessing the quality of research. *British Medical Journal, 328*, 39-42.

Greer, N., Mosser, G., Logan, G., & Halaas, G. W. (2000). A practical approach to evidence grading. *Journal on Quality Improvement, 26*(12), 700-712.

Gyorkos, T. W., Tannenbaum, T. N., Abrahamowicz, M., Oxman, A. D., Scott, E. A., Millson, M. E., et al. (1994). An approach to the development of practice guidelines for community health interventions. *Canadian Journal of Public Health, 85*(Supplement I), S8-S13.

Hadorn, D. C., Baker, D., Hodges, J. S., & Hicks, N. (1996). Rating the quality of evidence for clinical practice guidelines. *Journal of Clinical Epidemiology, 49(7*), 749-754.

Harden, A., Garcia, J., Oliver, S., Rees, R., Shepherd, J., Brunton, G., & Oakley, A. (2004). Applying systematic review methods to studies of people's views: An example from public health research. *Journal of Epidemiology and Community Health, 58*(9), 794-800.

Hernández-Avila, M., Garrido, F. y Salazar-Martínez, E. (2000). Sesgos en estudios epidemiológicos. *Salud Pública de México, 42*(5), 438-446.

Higgins, J. P. T., & Green, S. (2008). *Cochrane handbook for systematic reviews of interventions*. New York: John Wiley & Sons Inc.

Jüni, P., Altman, D. G., & Egger, M. (2001). Systematic reviews in health care: Assessing the quality of controlled clinical trials. *British Medical Journal 323*, 42-46.

Jüni, P., Witschi, A., Bloch, R., & Egger, M. (1999). The hazards of scoring the quality of clinical trials for meta-analysis. *Journal of the American Medical Association, 282*(11), 1054-1060.

Khan, K.S., ter Riet, G., Glanville, J., Sowden, A.J., & Kleijnen, J. (2001). *Undertaking systematic reviews of research on effectiveness. CRD's guidance for those carrying out or commissioning reviews*. CRD Report Number 4 (2nd Edition). Publications Office, NHS Centre for Reviews and Dissemination, University of York. Retrieved April 4, 2009, from http://www.york.ac.uk/inst/crd/pdf/crdreport4_complete.pdf

Kopec, J.A., & Esdaile, J.M. (1990). Bias in case-control studies. A review. *Journal of Epidemiology and Community Health, 44*(3), 179-186.

Kreulen, C. M., Creugers, N. H., & Meijering, A. C. (1998). Meta-analysis of anterior veneer restorations in clinical studies. *Journal of Dentistry, 26(*4), 345-353.

Krogh, C. L. (1985). A checklist system for critical review of medical literature. *Medical Education, 19*, 392-395.

Kwakkel, G., Wagenaar, R. C., Kollen, B. J., & Lankhorst, G. J. (1996). Predicting disability in stroke-a critical review of the literature. *Age and Ageing, 25(*6), 479-489.

Laupacis, A., Wells, G., Richardson, W. S., & Tugwell, P. (1994). Users' guides to the medical literature. V. How to use an article about prognosis. *Journal of the American Medical Association, 272(*3), 234-237.

Levine, M., Walter, S., Lee, H., Haines, T., Holbrook, A., & Moyer, V. (1994). Users' guides to the medical literature. IV. How to use an article about harm. *Journal of the American Medical Association, 271*(20), 1615-1619.

Lichtenstein, M. J., Mulrow, C. D., & Elwood, P. C. (1987). Guidelines for reading case-control studies. *Journal of Chronic Diseases, 40(*9), 893-903.

Liddle, J., Williamson, M., & Irwig, L. (1996). *Method for evaluating research guideline evidence (MERGE)*. Sydney: NSW Health Department.

Littenberg, B., Weinstein, L. P., McCarren, M., Mead, T., Swiontkowski, M. F., Rudicel, S. A., et al. (1998). Closed fractures of the tibial shaft. *The Journal of Bone and Joint Surgery, 80*(2), 174-183.

Loney, P. L., & Stratford, P. W.(1999). The prevalence of low back pain in adults: A methodological review of the literature. *Physical Therapy, 79(*4), 384-396.

López de Argumedo, M., Reviriego, E., Andrío, E., Rico, R., Sobradillo, N., & Hurtado de Saracho, I. (2006). *Revisión externa y validación de instrumentos metodológicos para la lectura crítica y la síntesis de la evidencia científica*. Vitoria-Gasteiz: Osteba-Servicio de Evaluación de Tecnologías Sanitarias. Departamento de Sanidad. Gobierno Vasco.

Mann, C. J. (2003). Observational research methods. Research design II: Cohort, cross sectional, and case-control studies. *Emergency Medicine Journal, 20*(1), 54-60.

Martin, J. L. R., Pérez, V., Sacristán, M., & Álvarez, E. (2005). Is grey literature essential for a better control of publication bias in psychiatry? An example from three meta-analyses of schizophrenia. *European Psychiatry, 20*(8), 550-553.

Margetts, B. M., Thompson, R. L., Key, T., Durry, S., Nelson, M., Bingham, S., et al. (1995). Development of a scoring system to judge the scientific quality of information from case-control and cohort studies of nutrition and disease. *Nutrition and Cancer, 24*(3), 231-239.

Margetts, B. M., Vorster, H. H., & Venter, C. S. (2002). Evidence-based nutrition - review of nutritional epidemiological studies. *South African Journal of Clinical Nutrition, 15*, 68-73.

New Zealand Guidelines Group (2001). *Handbook for the preparation of explicit evidence based clinical practice guidelines. Generic appraisal tool for epidemiology (GATE) methodology checklist*. Unpublished document. Retrieved January 9, 2010, from http://www.nzgg.org.nz/index.cfm?fuseaction=download&fusesubaction=template&libraryID=102

Nguyen, Q. V., Bezemer, P. D., Habets, L., & Prahl-Andersen, B. (1999). A systematic review of the relationship between overjet size and traumatic dental injuries. *European Journal of Orthodontics, 21*(5), 503-515.

Parker, S. (2006). *Guidelines for the critical appraisal of a paper*. Unpublished document. Retrieved January 9, 2010, from http://www.surgical-tutor.org.uk/default-home.htm?papers/appraisal.htm~right.

Pérez-Rıos, M., Ruano-Ravina, A., Etminan, M., & Takkouche, B. (2009). A meta-analysis on wood dust exposure and risk of asthma. *Allergy*, *65*, 467–473.

Rangel, S. J., Kelsey, J., Colby, C. E., Anderson, J. D., & Moss, R. L. (2003). Development of a quality assessment scale for retrospective clinical studies in pediatric surgery. *Journal of Pediatric Surgery, 38*(3), 390-396.

Reed, D. A., Cook, D. A., Beckman, T. J., Levine, R. B., Kern, D. E., & Wright, S. M. (2007). Association between funding and quality of pub-

lished medical education research. *Journal of the American Medical Association, 298*(9), 1002-1009.
Reisch, J. S., Tyson, J. E., & Mize, S. G. (1989). Aid to the evaluation of therapeutic studies. *Pediatrics, 84*(5), 815-827.
Sanderson, S., Tatt, I. D., & Higgins, J. P. (2007). Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: A systematic review and annotated bibliography. *International Journal of Epidemiology, 36*(3), 666-676.
Scottish Intercollegiate Guidelines Network. (2008). *SIGN 50: A guideline developer's handbook. Publication Nº 50.* Edinburgh: SIGN.
Shadish, W.R., Cook, T.D., & Campbell, D.T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston, MA: Houghton Mifflin.
Scottish Intercollegiate Guidelines Network. (2008). *SIGN 50: A guideline developer's handbook.* Edinburgh: SIGN.
Shaughnessy, J.J., Zechmeister, E.B., & Zechmeister, J.S. (2007). *Métodos de investigación en psicología,7a ed.* Madrid: McGraw-Hill.
Sica, G.T. (2006). Bias in Research Studies. *Radiology, 238* (3), 780-789.
Solomon, D. H., Bates, D. W., Panush, R. S., & Katz, J. N. (1997). Costs, outcomes, and patient satisfaction by provider type for patients with rheumatic and musculoskeletal conditions: A critical review of the literature and proposed methodologic standards. *Annals of Internal Medicine, 127(*1), 52-60.
Spitzer, W. O., Lawrence, V., Dales, R., Hill, G., Archer, M. C., Clark, P., Morgan, P.P. (1990). Links between passive smoking and disease: A best-evidence synthesis. *Clinical and Investigative Medicine, 13*(1), 17-42.
Steinberg, E. P., Eknoyan, G., Levin, N. W., Eschbach, J. W., Golper, T. A., Owen, W. F., et al. (2000). Methods used to evaluate the quality of evidence underlying the national kidney foundation-Dialysis outcomes quality initiative clinical practice guidelines: Description, findings, and implications. *American Journal of Kidney Diseases, 36(*1), 1-9.
Stock, S. R. (1991). Workplace ergonomic factors and the development of musculoskeletal disorders of the neck and upper limbs: A meta-analysis. *American Journal of Industrial Medicine, 19*(1), 87-107.
Streiner, D. L., Norman, G. R., & Fulton, C. (1991). Health measurement scales: a practical guide to their development and use. *International Journal of Rehabilitation Research*, *14*(4), 364.
The Joanna Briggs Institute (2008). *Joanna briggs institute reviewers' manual.* Unpublished document. Retrieved January 9, 2010, from http://www.joann riggs.edu.au/pdf/JBIReviewManual_CiP11449.pdf.
Tseng, T. Y., Breau, R. H., Fesperman, S. F., Vieweg, J., & Dahm, P. (2008). Evaluating the evidence: the methodological and reporting quality of comparative observational studies of surgical interventions in urological publications. *British Journal of Urology International,103,* 1026-1031.
Valentine, J.C., & Cooper, H. (2008) A systematic and transparent approach for assessing the methodological quality of intervention effectiveness research: the Study Design and Implementation Assessment Device (Study DIAD). *Psychological Methods, 13*(2), 130-149.
van der Windt, D. A., Thomas, E., Pope, D. P., de Winter, A. F., Macfarlane, G. J., Bouter, L. M., et al. (2000). Occupational risk factors for shoulder pain: A systematic review. *Occupational and Environmental Medicine, 57*(7), 433-442.
Vandenbroucke, J. P., von Elm, E., Altman, D. G., Gøtzsche, P. C., Mulrow, C. D., Pocock, S. J., Poole, C., Schlesselman, J.J., & Egger, M. (2007). Strengthening the reporting of observational studies in epidemiology (STROBE): Explanation and elaboration. *Epidemiology, 18*(6), 805-835.
Vitali, S. H., & Randolph, A. G. (2005). Assessing the quality of case-control association studies on the genetic basis of sepsis. *Pediatric Critical Care Medicine, 6(*3), S74-S77.
Weightman, A. L., Mann, M. K., Sander, L., & Turley, R. L. (2004). Questions to assist with the critical appraisal of an observational study e.g. cohort, case-control, cross-sectional. *Health Evidence Bulletins – Wales, 2009.*
Wells, G. A., et al. *The newcastle-ottawa scale (NOS) for assessing the quality of nonrandomized studies in meta-analyses.*, Unpublished document. Retrieved January 9, 2010, from http://www.ohri.ca/programs/clinical_epidemiology/oxford.htm.
Wells, K., & Littell, J. H. (2009). Study quality assessment in systematic reviews of research on intervention effects. *Research on Social Work Practice, 19*(1), 52-62.
Welsh Child Protection Systematic Review Group (2006). *Evidence sheet: Child prOtection neurological injuries CONI) critical appraisal forms. university of wales.* Unpublished document. Retrieved April 4, 2009, from http://www.core-info.cardiff.ac.uk.
West, S., King, V., Carey, T. S., Lohr, K. N., McKoy, N., Sutton, S. F., & Lux, L. (2002). *Systems to rate the strength of scientific evidence* (Evidence Report/Technology Assessment No. 47. AHRQ Publication No. 02-E016). Rockville, MD: Agency for Healthcare Research and Quality.
Wong, W. C., Cheung, C. S., & Hart, G. J. (2008). Development of a quality assessment tool for systematic reviews of observational studies QATSO) of HIV prevalence in men having sex with men and associated risk behaviours. *Emerging Themes in Epidemiology, 5*, 23-26.
Zaza, S., Wright-De Agüero, L. K., Briss, P. A., Truman, B. I., Hopkins, D. P., Hennessy, M. H., Pappaioanou, M. (2000). Data collection instrument and procedure for systematic reviews in the Guide to Community Preventive Services. Task force on community preventive services. *American Journal of Preventive Medicine, 18(*1S), 44-74.
Zola, P., Volpe, T., Castelli, G., Sismondi, P., Nicolucci, A., Parazzini, F., et al. (1989). Is the published literature a reliable guide for deciding between alternative treatments for patients with early cervical cancer? *International Journal of Radiation Oncology, Biology, Physics, 16(*3), 785-797.

(Article received: 20-12-2010, reviewed: 26-09-2011, accepted: 01-10-2011)