Doctoral Thesis

Q-Coh

A tool to assess the methodological quality of cohort studies

Alexander Jarde

Directed by Josep Maria Losilla and Jaume Vives

Estudis de Doctorat en Psicologia de la Salut i Psicologia de l'Esport
Departament de Psicobiologia i Metodologia de les Ciències de la Salut
Facultat de psicologia
Universitat Autònoma de Barcelona
2013

Index

1 Prologue

My initial intention when I started my research was to develop a questionnaire to assess psychosocial risk factors at work. As is widely recommended, my first task had to be making a systematic review of the topic. However, as I started to learn how to perform a systematic review I noticed that I needed a tool to assess the quality of the located studies, mostly of non-randomized nature. On my search for this tool I found an article by Sanderson, Tatt, and Higgins (2007), a systematic review of quality assessment tools for observational studies, which I expected would help me choose the best quality assessment tool. However, the authors concluded that there was “no single obvious candidate tool for assessing quality of observational epidemiological studies”. Therefore, to develop a questionnaire to assess psychosocial risk factors at work I had to do a systematic review first, which required me to have a quality assessment tool for observational studies, of which none of the existing ones was good enough. So I shifted my goal to the development of such a tool, embarking in a stirring journey into the roots of research methodology and up the branches of meta-analysis and evidence based science.

2 Introduction

2.1 Narrative vs. Systematic Reviews

For decades there have been authors, usually experts in the correspondent area, who have collected and summarized the published studies on certain topics. However, these so-called narrative reviews were increasingly criticized for their subjectivity and lack of transparency. In contrast, a more objective and transparent approach appeared: systematic reviews (Detsky, Naylor, O'Rourke, McGeer, & L'Abbé, 1992).

A systematic review is defined as "the application of scientific strategies that limit bias by the systematic assembly, critical appraisal and synthesis of all relevant studies on a specific topic" (Cook, Sackett, & Spitzer, 1995; Crowther & Cook, 2007). Therefore, there are several differences between systematic and narrative reviews. In narrative reviews the search strategy is typically not provided, while it is explicitly reported in a systematic review. A systematic review and involves a comprehensive search of many bibliographic databases as well as the so-called grey literature, which is literature not identifiable through a traditional index or database (McKimmie & Szurmak, 2002), such as personal communications or reports (Martin, Pérez, Sacristán, & Álvarez, 2005). The selection of the primary studies is based on explicit criteria which are uniformly applied in systematic reviews, while usually unspecified criteria are used in narrative ones. The review of the primary studies is typically done using a data extraction form in systematic reviews, including some assessment of the quality of the study. However, the article review process is variable in narrative reviews and the quality is usually not assessed. (Manchikanti, 2008). Thus, the major difference between a systematic review and narrative reviews is that a systematic review attempts to minimize bias by the comprehensiveness and reproducibility of the search and selection of articles for review and by assessing the methodological quality of the studies (Khan, Daya, & Jadad, 1996).

2.2 Meta-analysis

Once the literature has been reviewed it is usually interesting to quantitatively summarize the results of the reviewed studies (primary studies). Throughout research history there have been several methods for this, but the most commonly used statistical technique nowadays is meta-analysis: the statistical combination of results from two or more separate studies (Higgins & Green, 2011). So, although there may be systematic reviews without a meta-analysis, generally it is the final step in such reviews. To do so, it is necessary to first quantify the results of each primary study using a common effect size index. Once this is done it is possible to apply statistical analysis techniques to combine these effects sizes. Usually a systematic review that applies meta-analytic techniques is called just a meta-analysis (Sánchez-Meca & Botella, 2010).

2.3 Advantages of systematic reviews and meta-analyses

Systematic reviews and meta-analyses provide clinicians, researchers, policy-makers, and patients with a synthesis of an unmanageable and exponentially increasing number of manuscripts by linking and correlating huge amounts of information with identification of beneficial or harmful interventions (Manchikanti, Benyamin, Helm, & Hirsch, 2009a; Martin et al., 2005). In addition, systematic reviews and meta-analyses increase the power of the investigations by combining the information from individual studies so that its overall sample size is greater than that of any of the primary studies; and they limit bias and improve the reliability and accuracy of recommendations because of their formalized and thorough method of investigation. Furthermore, systematic reviews may help determine whether the results are consistent from study to study and to generalize the results; and may reduce the delay between publication of research findings and the implementation of new effective treatment strategies (Petrie, Bulman, & Osborn, 2003).

Obviously, systematic reviews and meta-analyses have some limitations, too, which are related to the heterogeneity of the primary studies, publication bias, difficulties finding a common effect size index, and the quality of the primary study, among others (Marín-Martínez, Sánchez-Meca, Huedo, & Fernández, 2007). This last limitation is the one addressed in this doctoral thesis.

2.4 Quality of the primary studies

Given the benefits of performing a systematic review, it is not surprising that there has been an explosion of systematic reviews and meta-analyses in the last decade (Manchikanti et al., 2009a). However, several studies have shown that the quality of systematic reviews is highly variable (McAlister et al., 1999). Among other methodological problems analyzed, it has been shown that the assessment of the quality of the primary studies - which is a defining difference between systematic reviews and narrative reviews – is lacking in a high percentage of published systematic reviews (Jadad, 2000; Moher et al., 1999; Petticrew, Song, Wilson, & Wright, 1999).

This is an important point, since the results of the systematic review and meta-analysis may be severely affected by the quality of the primary studies (“garbage in, garbage out”; Detsky et al., 1992; Jadad, Moher, & Klassen, 1998; Jüni, Altman, & Egger, 2001). So, if the less rigorous studies are biased towards –for example- overestimating an intervention’s effectiveness, the results of the meta-analysis will be biased (false-positive conclusions).

However, the assessment of the quality of the primary studies is not as easy as it may seem, and some authors even consider the assessment of the quality of the primary studies as uninformative or worse: a source of bias by itself (Greenland, 1994; Moher, Jadad, & Tugwell, 1996). In fact, in a study where Jüni, Witschi, Bloch, and Egger (1999) compared multiple meta-analyses performed with the best primary studies considered by different quality assessment tools, the results showed that the conclusion of a review could change significantly depending on the tool used to assess the quality of the primary studies. However, in much the same way an inappropriate statistical test can lead to invalid conclusions, we think that those results just highlight the importance of the tool used to measure quality. The fact that there are still imperfect quality assessment tools does not mean that the incorporation of quality assessment into systematic reviews should not be done, but that more efforts have to be made to develop a valid and reliable quality assessment tool using rigorous criteria.

2.5 Observational studies in systematic reviews

Generally systematic reviews and meta-analyses are performed on randomized trials, as they are usually considered the gold standard in research (Abel & Koch, 1999; Byar et al., 1976; Feinstein, 1984). The major advantage is given by its defining trait –randomization- since it minimizes selection bias by controlling for known and unknown factors (confounders) that may affect the outcomes (Manchikanti, Hirsch, & Smith, 2008). However, a randomized trial is often unethical (e.g., randomization to a probable harmful exposure, or randomization to a drug that has proven benefits but uncertain side effects), inefficient (long-term and rare outcomes), or not feasible to do, therefore requiring observational research (Black, 1996; Greene, 2000; Higgins & Green, 2011, Mann, 2003), in which the researcher has no influence in the allocation of the exposure variable and only 'observes' the outcomes (Mann, 2003; West et al., 2002). Observational studies then provide the only source, or a large component, of relevant evidence (Thompson et al., 2010). Actually, most questions in medical research are investigated in observational studies (Funai, Rosenbush, Lee, & Del Priore, 2001; Scales, Norris, Peterson, Preminger, & Dahm, 2005; Vandenbroucke et al., 2007, von Elm et al., 2007). So, although less numerous than systematic reviews and meta-analyses of RCT (Detsky et al., 1992; Manchikanti, Datta, Smith, & Hirsch, 2009b), systematic reviews are also performed on observational studies. In fact, Hartz, Benson, Glaser, Bentler, and Bhandari (2003) suggested that the "review of several comparable observational studies may help evaluate treatment, identify patient types most likely to benefit from a given treatment, and provide information about study features that can improve the design of subsequent observational or randomized controlled trials".

2.6 Combining evidence from observational studies and controlled trials

It is debated what to do when the inclusion of both randomized and observational studies is available for a certain topic. An extreme opinion would state that no observational study should be included in the meta-analysis if there is data available from any randomized trials. However, this seems far too restrictive if only a few randomized studies are available, and even more if the quality of those few randomized trials is low, whereas the available observational studies are of high quality (Shrier et al., 2007). The 'results of observational studies sometimes, but not always, differ from results of randomized studies of the same intervention' (Deeks et al., 2003).Some studies comparing randomized controlled trials with different observational studies showed different results depending on the research design (Shikata, Nakayama, Noguchi, Taji, & Yamagishi, 2006), although other studies found little evidence that estimates of treatment effects in observational studies were different from those found in RCT (Benson & Hartz, 2000). However, these results may be confounded by the quality of the primary studies, since several authors have concluded that there were no significant differences between the results of RCT and those of well-designed observational studies (Concato, Shah, & Horwitz, 2000; MacLehose et al., 2000).

Despite there is some overlap between overall research design and internal validity (Wells & Littell, 2009), equaling research design with quality would be a mistake. For example, one of the main limitations ascribed to observational studies is confounding due to selection bias, which is intended to be reduced in RCT by randomizing the allocation of the subjects to each group. On one hand, several authors have questioned the concept that random allocation of the subjects to either experimental or controlled groups is that reliable (Kane, 1997; Shrier et al., 2007). On the other hand, selection bias can be strongly reduced in a carefully designed observational study (Shrier et al., 2007) and the plausibility of an unmeasured confounder can be measured (Groenwold, Nelson, Nichol, Hoes, & Hak, 2010; McMahon, 2003; Normand, 2005; Rosenbaum, 1991). Therefore, what matters is whether those to be compared are similar with respect to potential confounding factors, not whether it is accomplished through randomization (Bluhm, 2009).

In a similar way, many advantages routinely ascribed to the RCT design can be achieved through careful observational design, hence observational studies should not be excluded a priori' (Shrier et al., 2007). After all, non-randomized studies are a necessary part of the evidence base for practice, both because they are able to provide information about a larger and more diverse population of patients, and because they are more likely to follow patients at outcome over long periods of time (Manchikanti et al., 2009b). Therefore, RCT and observational studies should not be considered mutually exclusive, but complementary.

2.7 Quality assessment tools for observational studies

If the assessment of methodological quality is an essential part of a systematic review, more so for observational studies (Manchikanti et al., 2009b), since they are more prone to bias (Higgins & Green, 2011; Manchikanti, Singh, Smith, & Hirsch, 2009c).

When we started this research three systematic reviews of tools to assess observational studies (among others) had been published. The first of them, by West et al. (2002), considered five key domains and found six tools that covered at least four of them. In the review by Deeks et al. (2003) the selection was based on six domains and four key items, ending up with six tools, also. The overlapping of the tools highlighted by these reviews was of 50% (three tools). Finally, the most recent review at the beginning of this research was one by Sanderson et al. (2007), which focused explicitly on the three main types of observational designs: cohort, case-controls, and cross-sectional (Vandenbroucke et al., 2007). Their conclusion was that there is no single obvious candidate tool for assessing the quality of observational studies. Despite their discrepancies, however, there is one important aspect that all previous systematic reviews agree on: most of the existing tools lack a rigorous development procedure following standard psychometric techniques.

2.8 Steps in the development of any measurement tool

Considering that a quality assessment tool is still a measurement instrument, it has to be developed following standardized procedures. The American Psychological Association (APA), jointly with the American Educational Research Association (AERA) and the National Council on Measurement in Education (NCME) developed the Standards for Educational and Psychological Testing (APA, NCME, & AERA, 1999). Following these standards, there are four phases that all tests have to undergo. First, the purpose of the test and the scope of the construct to be measured have to be delineated. Second, the test specifications have to be developed. Third, the items and scoring guides and procedures have to be developed, tested, evaluated and selected. And fourth, the test has to be assembled and evaluated for its operational use.

2.8.1 Purpose of the test and construct to be measured.

The first step in the development of a measurement tool is to extend the purpose and the construct being considered into a framework for the test that describes the extent of the domain, or the scope of the construct to be measured (APA et al., 1999).

Purpose. The existing quality assessment tools were not all developed with the same purpose. Actually, the development of the tool with the specific purpose of being used as a generic tool for systematic reviews and meta-analysis is not the most frequent case. More often, these tools are developed for a single use in a specific context (e.g. developed for a specific systematic review or meta-analysis), or to be used to aid in the critical appraisal of studies (Sanderson et al., 2007).

Construct to be measured. There have been several approaches to the assessment of the quality of primary studies. Wells and Littell (2009) listed six of them: publication status, reporting quality, design hierarchies, design features, the validity framework, and the risk of bias framework.

Using the publication status of a study, relying on the peer review process, may be misleading, since publication decisions may be affected by factors other than study quality (Wells & Littell, 2009). Another approach has focused on the reporting quality. However, although reporting quality and quality of the study may be related, a clear distinction should be drawn between what is reported that was done and what was actually done. This has been increasingly clear with the appearance of communication guidelines like the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement (Vandenbroucke et al., 2007). However, in a study examining the uses and misuses of the STROBE statement, Costa, Cevallos, Altman, Rutjes, and Egger (2011) found that half of the systematic reviews they located inappropriately used STROBE as a tool to assess the study quality.

Another approach to the assessment of the quality of primary studies has focused on the study design, leading to the proposal of design hierarchies (e.g. Harbour & Miller, 2001). Typically, meta-analyses are placed at the top of the hierarchy, followed by randomized controlled trials and quasi-experimental designs. Beneath are observational studies, and non-experimental studies. However, as already discussed in the introduction, the potential internal validity of higher placed designs should not be taken for granted, since such a study may still be of low quality if it is poorly executed.

A fourth approach to the assessment of the quality of primary studies has focused on features of the study design and implementation. In this approach scholars assess if certain features related to the study quality have been performed, instead of taking them for granted by the design the study is labeled with.

In a fifth approach to the assessment of the quality of primary studies, the concept of study quality is equated with “the relative absence of threats to the validity of an intervention” (Wells & Littell, 2009). Therefore, this approach makes a step further than the previous approach by inferring an assessment of the validity (or absence of risk of bias) of the studies based on their design features. In this sense, Dreier, Borutta, Stahmeyer, Krauth, and Walter (2010) argue that quality assessment tools should focus on internal validity and, if external validity is assessed, it should be done separately from internal validity.

Finally, and in contrast with the previous mentioned approach, the risk of bias framework focuses not on the validity of the intervention, but on the threats to this validity, or risks of bias. From this point of view, separating the ‘better’ from the ‘poorer’ quality studies introduces a rather arbitrary dichotomy and essentially disregards any biases in the ‘better’ studies, and assumes that the ‘worse’ studies are totally non-informative (Thompson et al., 2010). Instead, the potential biases should be quantified, both their magnitude as their direction.

Framework. When defining the tool’s framework, one important step is to clearly define the type of studies it will be applicable to. Just mentioning the design label is not enough, though, since authors and databases of different fields use a variety of terminology. As an example, ‘longitudinal study’, ‘follow-up study’, ‘prospective study’ and ‘cohort study’ are closely related terms and are commonly used as synonyms (Vandenbroucke et al., 2007). This might not be surprising considering the fact that the definitions and relations between them are not consistent along different reference sources:

For instance, in psychology in Spanish-speaking countries a commonly used terminology is the one proposed by Montero and León (2007). These authors do not use the terms ‘follow-up study’ nor ‘cohort study’. Instead, they call a study were the exposure (or independent variable) is not manipulated and is registered before the response (or dependent variable) ‘prospective study’. In their terminology, a longitudinal study is one kind of descriptive study of populations that uses survey research.

On the other hand, looking up PsycInfo’s Thesaurus (PsycInfo is a leading database in psychology), one can find, under the term ‘Experimental Design’, the terms ‘Longitudinal studies’, ‘Follow-up Studies’ and ‘Cohort Analysis’ (among others). ‘Prospective Studies’ are nested under ‘Longitudinal studies’.

Other databases that use a thesaurus are Medline and the Cochrane Library, which is in both cases the Medical Subject Heading (MeSH). Here, the terms ‘Longitudinal Studies’, ‘Follow-up Studies’, and ‘Prospective Studies’ are all considered ‘Cohort Studies’. But depending on the tree (there are three trees were the term ‘Cohort Studies’ appear) the latter two are considered as subtypes of the first one. So, longitudinal, follow-up, and prospective designs are indeed cohort studies, but not all cohort studies have a longitudinal, follow-up, or prospective design.

Another approach to the problem of the labeling of study designs is the development of classification tools to identify study designs. Hartling et al. (2010) recently reviewed 23 such tools (including two of the Cochrane Non-Randomized Studies Methods Group) and developed and tested a new one based on the most preferred one (the criteria used to make this selection were ease of use, unique classification for each study design, ambiguity and comprehensiveness, among others). The resulting algorithm and glossary has no design labeled as ‘longitudinal study’ or ‘follow-up study’. ‘Prospective studies’ are considered one type of cohort studies (prospective cohort studies), together with retrospective cohort studies and non-concurrent cohort studies.

Given this variety and inconsistency (both by authors’ use and bibliographic database indexing), the Cochrane Non-Randomized Studies Methods Group advises those review authors interested in including non-randomized studies in their review not to rely on design labels, but to use explicit study design features (Higgins & Green, 2011).

2.8.2 Test specifications.

After deciding the content the test is pretending to measure and in which framework, the next step is to establish the test specifications. These delineate the format of items, tasks, or questions; the response format or conditions for responding; and the type of scoring procedures. All subsequent test development activities are guided by the test specifications (APA et al., 1999). Regarding quality assessment tools there are some characteristics that are nowadays out of doubt, although there are other aspects that are less clear.

Psychometric properties. A variety of indexes have been applied to assess the psychometric properties of quality assessment tools. Regarding the reliability, two indexes are usually reported: the test-retest (or intra-rater) reliability and the inter-rater reliability. In the context of systematic reviews and especially of meta-analyses it is common that several of the steps are done independently by two authors, and the assessment of the quality of the primary studies is no exception to this (Sánchez-Meca, 2010). Therefore, a good inter-rater reliability should be a primary objective when developing a quality assessment tool.

Two types of discrepancies may appear when different raters assess the quality of a primary study: discrepancies due to characteristics missed by one of the raters (e.g. one of them might have overlooked the sentence where it is said that outcome assessors were blinded), and discrepancies due to different judgments of the raters. The first type is easily solved, and when no consensus is found in the second type of discrepancies, usually a third author breaks the deadlock. However, this raises another question regarding the inter-rater reliability: would another research group applying the same quality assessment tool to the same study arrive to the same conclusion? This inter-groups reliability, which does not include the discrepancies due to errors or overlooked information, would be even more important than the inter-rater reliability, since it could determine the replicability of the meta-analysis' results.

Regarding the validity of a quality assessment tool, two indexes are usually reported: internal consistency and criterion validity. However, internal consistency - usually measured with Cronbach's alpha (Cronbach, 1951) - analyzes the correlation of different items measuring the same construct, which is not the case in quality assessment tools. For example, items assessing the reliability of the outcome and items assessing the control for confounding factors are measuring different things.

Criterion, concurrent or correlational validity, basically compare the results of the tool with another measure of the same construct. Ideally this would be a gold standard, but there is none when assessing the quality of research. Therefore, researchers who have developed quality assessment tools have compared their results with the criteria of experts (e.g. Reed et al., 2007), with the results of another quality assessment tool (e.g. Downs & Black, 1998; Cho & Bero, 1994), with a global assessment made by the same raters (e.g. Downs & Black, 1998), or with indexes like citation rates or impact factor of the journal (e.g. Reed et al., 2007).

Checklist vs. Scales. Although some discrepancies can be found in when to consider a tool as a checklist or a scale, the use of a summary score (which is probably the most defining trait of a scale, although it can be obtained for several checklist-type tools) is highly criticized (Deeks et al., 2003), considering that "they do not give a transparent estimation of the degree of bias" (Shamliyan et al., 2010). Effectively, summary scores neglect the information of individual items, and there is no empirical basis for the different weights that is (explicitly or implicitly) given to each item (Dreier et al., 2010). Furthermore, there are studies concluding that summary scores do not reflect the studies' validity (Herbison, Haysmith, & Gillespie, 2006) and others showing that studies have different summary scores depending on the scale that is applied (Jüni et al., 1999). So, most authors advocate for using checklists.

Objective items. Another aspect where there is little discussion is that items should be as objective as possible and imprecise terms like “appropriateness” should be avoided. As Dreier et al. (2010) point out, items should ask for the presence of concrete aspects instead of the appropriateness of a proceeding, and they should be operationalized in a precise, clear, and detailed form. However, not much more importance has been given to the items’ wording, although this is a widely studied aspect in survey methodology. This may be so because it is probably assumed that if the items are too subjective, this will be reflected in a poor inter-rater agreement, so the attention is shifted to this coefficient (Dreier et al., 2010).

Instructions. Although the presence of instructions guiding and aiding the use of the quality assessment tool is not consistently requested, there are authors strongly advocating for it (e.g. Deeks et al., 2003; Dreier et al., 2010). Instructions allow to define the used terminology and to give longer and more complete explanations than could be summarized wording up a single item. Finally, instructions also can guide the tool’s users in their decision-making process.

Knowledge of the user. There are discrepancies about the methodological and statistical knowledge that the tool’s user should have. While some authors argue that the quality assessment tool should be applied by methodologists, since they can better identify if potential biases have occurred (West et al., 2002), others consider that a quality assessment tool should be objective enough to be applied by non-methodological experts (Dreier et al., 2010). In fact, expecting that there will always be a methodologist among research groups performing a systematic review and meta-analysis is probably not realistic, despite being desirable.

Measurement level. Although not often mentioned, some authors have noted that several hypothesis or outcomes can be studied in a single study, each with more or less susceptibility to bias. Therefore, it is said that quality assessment tools should focus not only on the quality of the study as a whole, but also at the hypothesis/outcome level. (Dreier et al., 2010; Shamliyan et al., 2010). Similarly, it may be interesting to assess the quality at the group’s level, so that the procedures and measures applied to the exposed and control groups are assessed independently.

Orientation. It is not clear if the base consideration of a study should be of good or bad quality. If the tool considers the starting point as of good quality, the items will have to be more oriented to identifying aspects that downgrade this quality. On the other side, if the tool considers the starting point of the studies as low (or as not granted), its items will have to be oriented to identifying aspects that increase the confidence in the study's quality. In this sense, we have not found much discussion about what orientation a tool should have.

Furthermore, this aspect gets especially important when items cannot be answered because of a lack of reported information. Should not reported measures be considered as not done? While it is true that not everything that is done in a study is usually reported, it may be argued that if something that improves the study's quality is not reported, it was probably not done. But if no evidence is given about if something is done, it may be risky to assume it was. Judging a study's quality by its appearance is very risky, too, especially considering that observational studies are already prone to bias, and that there is less tradition of rigor than in randomized controlled trials (Manchikanti et al., 2009c).

Scope of the tool. It is not clear how specific a tool should be. On one hand, omnibus tools applicable to randomized and non-randomized designs seem to be too unspecific for the different challenges each individual design has to face. On the other hand, a highly topic specific tool could be of lesser use when comparing different meta-analyses. Some tools are less oriented to a research design, but to study objectives, as could be intervention effectiveness. Dreier et al. (2010) suggest that tools should be generic enough to be applied to different study fields and if design-specific items are used, caution should be taken.

Evaluation of the quality. As mentioned above, the use of a summary score is highly criticized. Instead, quality should be based on each component or domain (Shamliyan et al., 2010). Deeks et al. (2003) recommend making an evaluation of the quality following a 'mixed-criteria' approach, requiring objective data related to the study design and then making an evaluation of the quality. The procedure of the evaluation of the quality should be standardized and be as transparent as possible (Dreier et al., 2010).

Direction and magnitude of bias. For all potential sources of bias, it is important to consider the likely magnitude and the likely direction of the bias. For example, if all methodological limitations of studies were expected to bias the results towards a lack of effect, and the evidence indicates that the intervention is effective, then it may be concluded that the intervention is effective even in the presence of these potential biases. (Higgins & Green, 2011). In addition, if the direction and magnitude of the expected biases is assessed, researchers could adjust for them in the meta-analysis (Thompson et al., 2010).

Feasibility. The length and application time of the quality assessment tools are related characteristics of the tool that are not usually addressed in the literature, and if it is, vague criteria are given. So, application time is said to have to be moderate (Dreier et al., 2010). In addition, the tool's interface has to be user-friendly in order to avoid errors due to users' fatigue or misunderstandings. However, this aspect is rarely mentioned explicitly (e.g. Sanderson et al., 2007).

2.8.3 Development, testing, evaluation and selection of items.

The next step in test development is to assemble items into a test. Usually, the test developer starts with an item pool that is larger than required, which are then selected for the test following the requirements of the test specifications. The clarity, lack of ambiguity, and content quality of the items is usually ascertained through item review procedures and pilot testing. In addition, when evaluating the quality of the items in the item pool and the test itself, test developers often conduct studies of differential item functioning. This is said to exist when test takers differ in their responses to an item. When differential item functioning is detected, test developers try to identify plausible explanations for the differences in the item content, item format, or scoring criteria; and then they may replace or revise them (APA et al., 1999).

Development of the items. Typically, the development of items for quality assessment tools is based, when stated, on previous similar tools and/or methodological literature review. In less frequent cases a more empirical procedure is used to develop the items, as for example Delphi procedures (Whiting, Rutjes, Reitsma, Bossuyt, & Kleijnen, 2003) or a bank of items (Viswanathan & Berkman, 2011).

Pilot testing and evaluation. Once the pool of items is ready, they are organized and put together into a pilot version of the tool, which is then applied in a relative small sample. This pilot testing gives information about how users respond to the items, about the behavior of the different items, and allows a first measure of the test's psychometric properties: validity and reliability.

Validity. Usually, the validity of the individual items is evaluated by a group of experts who give their opinion on the face validity and relevance of each item. When a pilot testing is done, the feedback given by the pilot testers is often also taken into account.

Reliability. Regarding the reliability, while it is not frequent that the inter-rater reliability of a quality assessment tool is given, it is even rarer to find tools where the inter-rater reliability is reported at the item's level. However, this information should not be neglected considering that the items of a quality assessment tool are relatively independent of one another.

2.8.4 Assembling of the test.

There are different ways a test can be assembled once the items have been decided. The most often used format is that of plain text, as it has been for a long time the only way it could be published in journals. However, nowadays the electronic versions of some journals allow authors to attach other types of files to their papers, opening the possibility to other formats for the quality assessment tools. The use of electronic versions and adaptations of quality assessment tools offer several advantages. First, it allows recording the responses given to each item and automatically including it into a database. Second, the instructions to each item are easily made accessible. And third and probably most important: the quality assessment tool can be made dynamic, which includes skipping not applicable items, offering feedback and summaries of previous answered items, checking for inconsistencies, etc.

2.8.5 Summary checklist.

The checklist on Table 1 summarizes the characteristics and test specifications that a quality assessment tool should have.

2.9 Objectives of this research line

As mentioned in the introduction, most of the research in health sciences is of observational nature. However, their inclusion in systematic reviews and meta-analyses is still scarce and under debate. One of the reasons for this paucity is the lack of a valid and reliable tool to assess the methodological quality of observational studies. Therefore, the main objective of this research line is the development of such a tool using rigorous development procedures. The steps to reach this goal define our specific objectives:

1. Update the latest systematic review of quality assessment tools.
2. Appraise the need for a new quality assessment tool for observational studies.
3. Develop a quality assessment tool for cohort studies.
4. Analyze the psychometric properties of the tool.
5. Develop a quality assessment tool for case-control studies.
6. Analyze the psychometric properties of the tool.
7. Develop a quality assessment tool for cross-sectional studies.
8. Analyze the psychometric properties of the tool.
9. Unify the different quality assessment tools into a single tool.

Table 1. Summary checklist with the characteristics and test specifications that a quality assessment tool should have.

1	□	Rigorous development process: - Definition of construct to be measured - Item generation - Pretest - Reliability and validity
2	□	High inter-rater reliability.
3	□	No quantitative summary score (checklist, not a scale)
4	□	Objective items. Avoid imprecise terms.
5	□	Items ask for the presence of concrete facts.
6	□	Instructions: - Define the used terminology - Complete explanation of each item - Guidance when making decisions
7	□	Objective enough to be applied by non-methodological experts.
8	□	Moderate application time.
9	□	Discern quality of reporting from study quality.
10	□	If external validity is assessed, it should be done separately from internal validity.
11	□	Hypothesis level taken into account.
12	□	Generic enough to be applicable to different study fields.
13	□	Quality based on domains.
14	□	'Mixed-criteria' approach.
15	□	Standardized evaluation procedure.
16	□	Transparent evaluation process.
17	□	Assessment of the magnitude and direction of bias.
18	□	User-friendly interface.

2.10 Specific objectives and structure of this doctoral thesis

Three articles are presented in this compendium of publications. The first one is a systematic review of the tools published so far to assess the methodological quality of observational studies, therefore achieving our first specific objective and addressing the second one. As the psychometric properties of these tools were mostly lacking, in our second study we selected the three tools that best cover our domains of methodological quality and analyzed their psychometric properties. The results made it difficult to recommend any of the analyzed tools without reservations, especially regarding their application to cohort and case-control studies. This covered our second specific objective. Considering our results in the second study, we developed a quality assessment tool for cohort studies, which is published in our third study. With this study we covered the third and fourth specific objective.

Although not all of our specific objectives have been covered in this thesis, the most important ones have. These are the ones justifying the need for this research line (objectives one and two) and a first proposal of a tool (objectives three and four) which shows its results and potential.

3 Articles that form this compendium

3.1 Methodological quality assessment tools of non-experimental studies: a systematic review (Article 1)

The first of our specific objectives was to update the latest systematic review of quality assessment tools. This was necessary in order to know what tools had been done so far. Previous systematic reviews had had differing results: 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) concluded that there is no single obvious candidate tool for assessing quality of non-experimental studies. The only important aspect that all previous systematic reviews agreed on was that most of the existing tools had not been developed using standard psychometric techniques.

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 (plus the first 300 links using Google's search engine), yielding a total of 74 tools that were included in the review.

In order to make some judgment about the contents addressed by the quality assessment tools found it was necessary to explore and define the domains underlying the concept of methodological quality. Therefore, 6 domains of quality were defined based on reporting guidelines, the established bibliography, and previous similar studies: Representativeness, Selection, Measurement, Data Collection, Statistics and Data Analysis, and Funding.

In addition, five desirable aspects related to the tool's development were defined: Some discussion about the concept of "quality", information about the item selection process, performance of a pilot study, reliability testing, and validity testing.

Data of each tool was extracted independently by two of the authors using a Microsoft Access database (the manual and screenshots of the data extraction form are attached as Annex 1 and Annex 2), with differences of opinion resolved by discussion or by the third author.

As expected, our results confirmed that most of the reviewed tools had not been developed using rigorous standard procedures, with only 5 of them covering at least four of the desirable aspects during their development.

Regarding the number of domains somehow addressed by the tools, only 11 had at least one item related to each of our domains (excluding Funding).

Although the tool by Downs and Black (1998) was in an upstanding position, we were reluctant to recommend this tool yet, since it had other flaws not systematically analyzed in the review.

The results of this systematic review were presented at the XI Congress of Methodology of the Social and Health Sciences in Málaga (2009) and the paper was accepted for publication in Anales de Psicología on October 2011.

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 Sanderson 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.

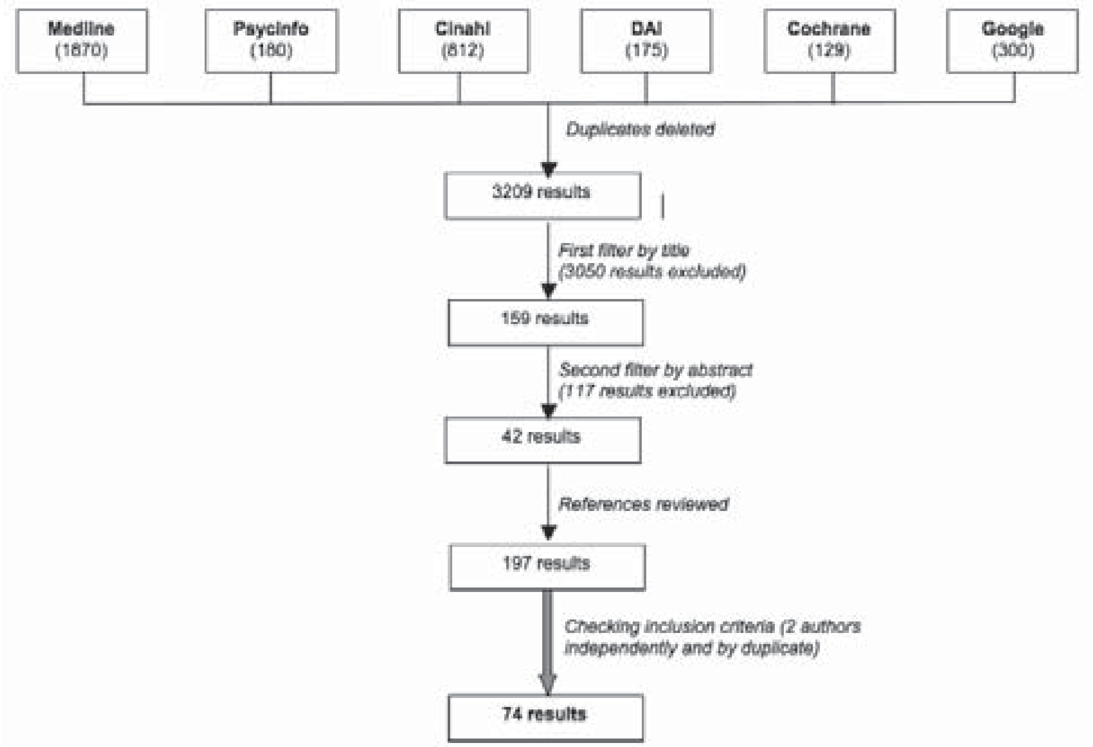


Figure 1. Search and selection process of all documents included in this review.

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 comparison 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 published 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)

3.2 Suitability of three different tools for the assessment of methodological quality in ex post facto studies (Article 2)

In the systematic review of quality assessment tools we had found 11 tools that had at least one item related to most of our domains of quality. However, although the validation of these tools had limitations or (as in most cases) was absent, we could not discard their usefulness (“the absence of evidence is not evidence of absence”). Therefore, in this second study we analyzed the psychometric properties of the three tools that best covered our domains of quality in order to confirm the need for a new tool (which was our second specific objective). To do so, we applied each of these tools to up to 30 studies of different designs (10 cohort, 10 case-control, and 10 cross-sectional studies).

For each research design and quality assessment tool six scores were recorded: one for each domain of quality by adding up the items related to each domain (the domain Funding was excluded of this procedure, since it was only considered in one tool), and one global score by adding up all items of the tool (regardless if they were related or not to any domain).

On one hand, to analyze if the tools identified the same strengths and weaknesses of the studies, it was not appropriate to focus on the global scores. So, although two tools could reach the same global score for a study, the strengths and weaknesses identified by each one could be different. Focusing on the agreement scores at the domain level, we actually were comparing groups of related items. Good agreements between tools indicated that they measure similar constructs, giving an indirect measure of concurrent validity.

On the other hand, to study the inter-rater agreement we focused on the global score of the complete tools, and separately for each research design. Good inter-rater agreements indicated that similar results should be expected for different raters.

The results showed different behaviors of the tools depending on the study design to which they were applied. While the inter-rater reliability of the three tools analyzed ranged from moderate to high for cross-sectional studies, for cohort or case-control studies only the tool by Downs and Black (1998) showed a moderate inter-rater agreement. However, while this was true for the global score, there was no consistent inter-rater reliability at the domain’s level.

Regarding the agreement between tools, despite analyzing it at the domains level (where a higher agreement should be expected) it was low in general.

The results of this study were presented at the IV European Congress of Methodology in Potsdam (2010) and the paper was accepted for publication in the International Journal of Clinical and Health Psychology on September 2011.

© International Journal of Clinical and Health Psychology

ISSN 1697-2600 print
ISSN 2174-0852 online
2012, Vol. 12, Nº 1, pp. 97-108

Suitability of three different tools for the assessment of methodological quality in ex post facto studies[1]

Alexander Jarde[2], Josep M. Losilla, and Jaume Vives
(*Universitat Autònoma de Barcelona, Spain*)

ABSTRACT. There is no clear candidate tool for assessing the methodological quality of *ex post facto* studies in systematic reviews and meta-analyses yet. Our purpose is to thoroughly analyze the psychometric properties of the three most comprehensive assessment tools of this kind published up to 2010. We selected these tools from a previous systematic review, and we applied each one to assess the quality of 10 prospective studies, 10 retrospective studies with quasi-control group, and 10 cross-sectional studies. Inter-rater reliability for the first two aforementioned research designs is moderate only for one of the selected tools, and moderate to high for all of them for cross-sectional studies. Agreement between tools is low in general, although the inferred aspects show that the tools have a relative good conceptual overlapping in most of the domains. According to these results we recommend two tools for assessing cross-sectional studies, but we consider that the tools applicable to prospective studies or retrospective studies with quasi-control group require further testing. The 30 concrete aspects that we have inferred from the items of the three analyzed tools can be used as starting point to develop a new tool of this kind.

KEYWORDS. *Ex post facto* studies. Quality assessment tools. Systematic reviews. Meta-analyses. Instrumental study.

[1] This research was supported by Grant PSI2010-16270 from the Spanish Ministry of Science and Innovation.

[2] Correspondence: Departament de Psicobiologia i de Metodologia de les Ciències de la Salut. Facultat de Psicologia. Campus de la Universitat Autònoma de Barcelona (UAB). 08193 Cerdanyola del Vallès. Barcelona. Spain. E-mail: A.Jarde@gmail.com

RESUMEN. No hay todavía un candidato claro a la hora de elegir una herramienta para valorar la calidad metodológica de estudios no experimentales en revisiones sistemáticas y meta-análisis. Nuestro propósito es analizar en profundidad las características psicométricas de las tres herramientas de evaluación de este tipo más comprensivas publicadas hasta el 2010. Seleccionamos estas herramientas de una revisión sistemática previa, y aplicamos cada una de ellas para valorar la calidad de 10 estudios prospectivos, 10 estudios retrospectivos con cuasi control y 10 estudios transversales. La fiabilidad entre jueces para los dos primeros diseños mencionados es moderada sólo en una de las herramientas seleccionadas, y moderada a alta en todas ellas para los estudios transversales. El acuerdo entre herramientas es en general bajo, pese a que los aspectos inferidos muestran que tienen un solapamiento conceptual relativamente bueno en la mayoría de las dimensiones. De acuerdo con estos resultados recomendamos dos herramientas para valorar estudios transversales, ya que consideramos que las herramientas aplicables a estudios prospectivos o retrospectivos con cuasi control requieren análisis adicionales. Los 30 aspectos concretos que hemos inferido de los ítems de las tres herramientas analizadas pueden usarse como punto de partida para desarrollar una nueva herramienta de este tipo.

PALABRAS CLAVE. Estudios *ex post facto*. Herramientas de evaluación de la calidad. Revisiones sistemáticas. Meta-análisis. Estudio instrumental.

It is very important to thoroughly appraise methodological quality of the primary studies when performing systematic reviews and meta-analyses, because if the primary studies are flawed, then the conclusions cannot be trusted (Jüni, Altman, and Egger, 2001; Jüni, Witschi, Bloch, and Egger, 1999; Valentine and Cooper, 2008). Therefore, studies have to be included/excluded or weighted according to their quality or probability of bias.

Although the inclusion of experiments in systematic reviews and meta-analyses is well established, the inclusion of non-experimental studies is still under debate, as they are more prone to certain biases (Shrier *et al*., 2007). However, these designs cannot be ignored, since they are often the most efficient ones to answer certain questions and may even be the only practicable method of studying certain problems. That is why a reliable assessment tool of their methodological quality is needed. Dozens of such tools have been proposed so far, but few of them are developed following standardized procedures (Carretero-Dios and Pérez, 2007) and there is no consensus on which tool is the most appropriate to evaluate *ex post facto* studies (Sanderson, Tatt, and Higgins, 2007; Wells and Littell, 2009).

On the other hand, there are widely accepted proposals about the reporting quality of *ex post facto* studies. Although the quality of the information that appears published has to be clearly separated from the methodological quality of a study, they are closely related. In this regard, the STrengthening the Reporting of OBservational studies in Epidemiology (STROBE) Statement (Vandenbroucke *et al*., 2007) is endorsed by a growing number of biomedical journals. It is a checklist that provides guidance to authors about how to improve the reporting of cohort, case-control, and cross-sectional studies. In the epidemiological tradition, these designs are usually referred to as «observational studies» because no intervention is carried out by the researcher. This is also the main characteristic that defines *ex post facto* studies in Montero and León's terminology (2007), which is used in this journal. In order to avoid terminology confusions, especially among habitual readers of epidemiological literature, it should be noted that in this paper the authors have used the methodological classification of research studies proposed by the International Journal of Clinical and Health Psychology (IJCHP) editors instead of that generally used in epidemiology and suggested by the STROBE statement. Therefore, we used «prospective» instead of «cohort» design, and «retrospective design with quasi-control group» instead of «case-control» design. For more detailed information about observational designs, we recommend the article by Mann (2003).

We conducted a systematic review of methodological quality assessment tools of prospective studies, retrospective studies with quasi-control group, and cross-sectional studies published up to 2010 (Jarde, Losilla, and Vives, in press). The search was done in Medline, Psycinfo, Cinahl, Dissertation Abstracts International, Cochrane Library, and in the World Wide Web using the Google search engine (http://www.google.com) to locate gray literature (Fernández-Ríos and Buela-Casal, 2009). The inclusion and exclusion criteria for 197 eligible documents were checked, identifying 74 tools. We also proposed six domains of methodological quality based on reporting standards (the STROBE statement by Vandenbroucke *et al*., 2007, and JARS of the American Psychological Association, 2010), previous similar reviews (Deeks *et al*., 2003; Sanderson *et al*., 2007; West *et al*., 2002), and well-established methodological literature. Based on these domains of quality, 11 tools were highlighted for having at least one item related to each domain (or each domain except Funding). The domains were defined as follows:

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.
2. Selection. The different groups of participants are comparable on all important characteristics except on the variables under study.
3. Measurement. The instruments used to collect the data are appropriate (valid and reliable).
4. Data collection. The comparability of the groups and the data quality are not affected by threats that may appear during data collection and management.
5. Statistics and data analysis. Confounding is controlled and missing values and losses to follow-up are properly treated in the statistical analysis.
6. Funding. The sources of funding and possible conflicts of interests have not influenced the study.

Our purpose is to analyze the psychometric properties of the quality assessment tools that best cover these domains of methodological quality in order to recommend the best subset for its use in systematic reviews and meta-analyses of *ex post facto* studies. First, the characteristics related to the usage of the tools when applied to studies with prospective, retrospective with quasi-control group, or cross-sectional research designs are analyzed. Second, the inter-rater reliability is analyzed, since this is a key element if the tool has to be applied across systematic reviews and meta-analyses. Third, agreement between tools is analyzed in order to see if they are measuring the same underlying constructs. And fourth, items related to the domains of quality are arranged with concrete aspects within each domain in order to study the theoretical overlap between them.

Method

Selection of the tools

After scoring each of the 11 tools highlighted in our previous systematic review according to how far they covered each domain of quality, only 3 tools covered all domains (or all except Funding) better than just superficially or indirectly: the one by Berra, Elorza-Ricart, Estrada, and Sánchez (2008), applicable to cross-sectional studies only; the one by Downs and Black (1998), which is applicable to randomized and non-randomized studies; and the one by Fowkes and Fulton (1991), designed to assess experimental designs, as well as prospective, retrospective with quasi-control group, and cross-sectional designs.

Procedure

The selected tools were applied (when possible) to 30 studies (10 studies with prospective design, 10 studies with retrospective design with quasi-control group, and 10 cross-sectional studies) independently by two of the authors (AJ and JV). As each tool uses a different scoring system, and in order to be able to compare them, we recoded the scores so that higher scoring represented better quality (starting at zero). For each research design and quality assessment tool we calculated six scores, one for each domain of quality by adding up the items related to each domain (the domain Funding was excluded of this procedure, since it was only considered in one tool), and one global score by adding up all items of the tool (regardless if they were related or not to any domain).

To study the inter-rater agreement we focused on the global score of the complete tools, and separately for each research design. There could be a good inter-rater agreement on several domains, but this does not necessarily imply a good agreement on the global score, since it is computed using all the items (not just those related to a domain of quality). Good inter-rater agreements indicate that similar results should be expected for different raters. We compared the indexes of agreement between raters computed for each tool using the parametric intraclass correlation coefficient (ICC) for the global scores (Shrout and Fleiss, 1979). For the domains scores we used the nonparametric Kendall tau-b correlation coefficient (Kendall, 1938) because the multivariate normality assumption of the ICC was not satisfied.

On the other hand, to analyze if the tools identify the same strengths and weaknesses of the studies, it is not appropriate to focus on the global scores. So, although two tools can reach the same global score for a study, the strengths and weaknesses identified by each one can be different. Focusing on the agreement scores at the domain level, we are actually comparing groups of related items. Good agreements between tools indicate that they measure similar constructs, giving an indirect measure of concurrent validity. To evaluate the agreement between tools we applied the correlation coefficient because ICC is not applicable given that the maximum score is different for each tool.

Additionally, we analyzed which aspects of the domains where assessed by each item in order to study the theoretical overlap between tools. To do so, each author classified the items of all tools into subcategories within each domain. Then the three drafts of items classification and subcategory labeling were discussed until consensus was reached.

Results

Characteristics of the selected tools and their usage

Berra *et al*. (2008). This tool was developed to assess cross-sectional studies only and is written in Spanish. It has 27 items and the authors took into consideration literature on strength of evidence, other existing tools, and the STROBE statement recommendations (Table 1 shows some example items). No further information about its development or reliability and validity is given, though. It took 18 minutes on average to apply this tool, and the mean number of not applicable items was three (11% of the tool's items).

Downs and Black (1998). A pilot version of this tool was developed based on epidemiologic principles, reviews of study designs and previous quality assessment tools for randomized controlled trials. An explicit definition of the concept of quality is not given, though. The definitive version of this tool resulted from the corrections after testing the pilot version. Several reliability scores are given: Internal consistency using the Kuder-Richardson formula (KR-20 = .89), the Spearman correlation coefficient for test-retest (r = .88), and inter-rater reliability (r = .75) for the total score when applied to randomized and nonrandomized studies. Reliability of the sub-scales when applied to nonrandomized studies ranged from 0 to .59. Validity was assessed by comparing the tool's score with a global score provided by the reviewers (r = .86). The tool has 27 items and is claimed to be applicable to both randomized and nonrandomized studies (Table 1 shows some example items). In fact several items make specific reference to prospective and retrospective with quasi-control group designs, but cross-sectional designs do not seem to be taken into account. It took 19 minutes on average to apply this tool. The mean number of not applicable items was seven on prospective studies, eight on retrospective studies with quasi-control group, and ten on cross-sectional studies, which is more than one third of the tool's items.

Fowkes and Fulton (1991). This tool is designed to assess experimental designs, as well as prospective, retrospective with quasi-control group, and cross-sectional designs. It has 22 items and, although the authors discuss what their tool does and does not assess, no more information about its development or regarding its reliability and validity scores is given (Table 1 shows some example items). It took 12 minutes on average to apply this tool. The mean number of not applicable items was seven on prospective studies, six on retrospective studies with quasi-control group, which is more than 25% of the tool's items; and nine on cross-sectional studies.

TABLE 1. Example items from each tool for each domain of quality.

Berra et al. (2008)	*Downs and Black (1998)*	*Fowkes and Fulton (1991)*
Representativeness		
4. The study population defined by the selection criteria contains an adequate spectrum of the population of interest.	11. Were the subjects asked to participate in the study representative of the entire population from which they were recruited?	2.4. (…) Any description of the study participants must be scrutinized in order to assess whether the sample was representative.
Selection		
2. The participants' inclusion and exclusion criteria are described, as well as the sources and methods of selection.	5. Are the distributions of principal confounders in each group of subjects to be compared clearly described?	3.3. Did the matching process seem to have been carried out correctly?
Measurement		
12. The main variables have an adequate conceptual (…) and operational definition (…).	20. Were the main outcome measures used accurate (valid and reliable)?	4.1. It is important to assess the validity of measurements made in a research study (…).
Data collection		
9. The same measurement strategies and techniques were used in all groups; the same variables were measured in all groups.	15. Was an attempt made to blind those measuring the main outcomes of the intervention?	6.1. Could there possibly be extraneous treatments which might have influenced the results?
Statistics and data analyses		
18. The main possible confounding factors were taken into account in the design and in the analysis.	26. Were losses of patients to follow-up taken into account?	6.5. Distorting influences may be minimized by some form of stratification or adjustment procedure in the analysis.

Note. Berra *et al*'s (2008) items were translated from Spanish.

Inter-rater agreement

Inter-rater agreement varied depending on the research design to which the tools were applied. So, when assessing cross-sectional studies, all three tools had moderate to high inter-rater agreement both when the global score and the different domain scores were considered. On the contrary, when prospective designs and retrospective designs with quasi-control group were addressed, the tool by Downs and Black (1998) had moderate inter-rater agreement, and Fowkes and Fulton (1991) had low agreements (see Table 2 for details).

TABLE 2. Inter-rater agreement for each tool and design considering the global score and each domain score.

Tool	*Design*	*Represen-tativeness*	*Selection*	*Measure-ment*	*Data collection*	*Statistics & Data analysis*	*Global*
Downs and	P	.509	.592*		.816*	.625*	.695**
Black	R	.214	.429	-.612		.241	.605
(1998)	CS	.809*	.901**			.816*	.853**
Fowkes and	P	.711**	.059	.16	.323	.464	.253
Fulton	R	.027	.162	.247	.086	.383	-.044
(1991)	CS	.676*	.659*	.830**	.554	.806**	.759**
Berra *et al.* (2008)	CS[a]	.875**	.623*	.912**	.845**	.694*	.842**

Note. Intraclass correlation coefficient was used for the global scores. Kendall's tau-b correlation coefficient and its significance test were used for the domain scores. This value could not be calculated (blank cells) when all studies had the same score. P = prospective design; R = retrospective design with quasi-control group; CS = cross-sectional design. [a]Berra *et al*.'s tool is only applicable to cross-sectional studies. * $p < .05$; ** $p < .01$.

Agreement between tools

Table 3 shows the agreement coefficients of each domain's score and of the global score. These are presented separately for each rater since agreement between tools varied greatly across them. As we are interested in the agreement between tools on the strengths and weaknesses of the assessed studies we will focus mainly on the agreement coefficients within each domain of quality. Using these comparisons a higher agreement should be expected, since it compares groups of related items. With this in mind, our results show that globally there is not much agreement among the tools, independently of the rater. There are some consistent agreements between some tools for certain domains, though, when cross-sectional studies are assessed. The tools by Berra *et al*. (2008) and Fowkes and Fulton (1991) have moderate to high agreement on the domains Representativeness and Selection. Downs and Black's (1998) tool has a good agreement with Berra *et al*.'s (2008) on the domain Statistics and Data Analysis, and with Fowkes and Fulton's (1991) on the domain Selection. Looking separately at each rater's coefficients we can see that, for one of the raters, there is a moderate to high agreement between Downs and Black's (1998) and Fowkes and Fulton's (1991) tools on the domain Representativeness across all three designs to which these tools are applied. On the other hand, in the second rater's data what catches the eye is the fact that all tools have a good agreement on all designs when the global scores are compared, but this is not reflected in agreements on the different domains. Finally, the agreement coefficients of some domains could not be calculated because all studies had the same score on them when the tool by Downs and Black (1998) was applied.

TABLE 3. Tool's agreement coefficients (and p values) on the global and domains' score for each design and rater.

Rater 1 (AJ)					
	Prospective studies	Retrospective studies[a]	Cross-sectional studies		
Domains of quality[b]	D&B-F&F	D&B-F&F	D&B-F&F	D&B-Berra	F&F-Berra
Global score	.708**	.364	.446	.496	.636*
Representativeness	.743*	.847**	.651*	.488	.796**
Selection	.294	.257	.806**	.684*	.532*
Measurement	.500	-.069	.205	-.085	.361
Data collection	.244		.566	-.254	.118
Statistics and Data Analysis	.467	.213	.733*	.816**	.359
Rater 2 (JV)					
	Prospective studies	Retrospective studies[a]	Cross-sectional studies		
Domains of Quality[b]	D&B-F&F	D&B-F&F	D&B-F&F	D&B-Berra	F&F-Berra
Global Score	.588*	.659**	.548*	.674**	.595*
Representativeness	.487	.396	.558	.621*	.737**
Selection	.471	.514	.619*	.459	.676*
Measurement	NA	.313			.394
Data collection	.361	.507			.147
Statistics and Data Analysis	.348	.522	.431	.600*	.239

Note. Kendall's tau-b correlation coefficient and its significance test were used. This value could not be calculated (blank cells) when all studies had the same score. D&B = Downs and Black (1998); F&F = Fowkes and Fulton (1991); Berra = Berra *et al*. (2008). [a]Retrospective studies with quasi-control group. [b]There are no agreement coefficients for the domain Funding, since it was only considered in one tool. * $p < .05$; ** $p < .01$.

Aspects covered by the tools' items

A total of 30 aspects were inferred from the tools' items that were related to each domain of quality. A little more than half of these aspects (16) were covered by at least two tools, but the other half (14) were aspects only considered by one single tool. Table 4 shows which aspects of each domain of quality are covered by the items of each tool (some aspects are assessed by several items). As some items are double-barreled or very broad they can be assessing several aspects at the same time.

TABLE 4. Which aspects of each domain of quality that are covered by which item of each tool.

Domains and Aspects	*D&B*	*F&F*	*Berra*
1. Representativeness			
Representativeness of situations	13		
Similar distribution of confounders in sample and population	12	2.1	4
Comparability between participants and non-respondents	12	2.5	6
Sampling procedure	11	2.2	2, 4
Sample size large enough to be representative		2.3	
2. Selection			
Inclusion and exclusion criteria	3	2.4, 3.1	2
Similar distribution of confounders in all groups	5, 21	3.2, 3.4, 6.4	7, 8, 18
Participants of different groups recruited in similar moments	22		
Matching process carried out correctly		3.3	
3. Measurement			
Valid measurement tools	20	4.1	13
Reliable measurement tools	20	4.2	13
Conceptual and operational definition of the main variables			12
Calibration and accuracy of instruments		4.4	
4. Data collection			
Study subjects blind	14	4.3	
Those collecting the data blind	15	4.3	
Compliance	19		
Contamination	19	6.2	
History and/or maturation		6.1	
Changes over time		6.3	
Recall bias		4.3	14
Interviewer bias		4.3	14
Same measurements in all groups			9
Quality control measures		4.4	
Comparability not affected by losses to follow-up	9	5.2, 5.3	10
Comparability not affected by missing data		5.4	
5. Statistics and data analysis			
Adjustment for confounding in the analyses	25	6.5	18, 21
Adjustment for incomplete data	26	5.1	17
Adjustment for time lengths	17		
6. Funding			
Source of funding mentioned			27
Consideration of conflicts of interest			27

Discussion

We have found three tools that cover all our domains (or all except Funding) more than just superficially or indirectly. The application time varies depending on the design of the assessed study and the tool used, ranging between 10 and 23 minutes on average. Inter-rater reliability of the three tools analyzed ranged from moderate to high for cross-sectional studies. For prospective studies or retrospective studies with quasi-control group only the tool by Downs and Black (1998) showed a moderate inter-rater agreement. Agreement between tools was low in general, despite analyzing it at the domains level where a higher agreement should be expected. The inferred aspects show that the tools have a relative good conceptual overlapping in most of the domains except in the domain Data collection. This finding may suggest that the low indexes of agreement between tools are more related with characteristics of the items or with the different coverage of the quality domains than with a different underlying construct of quality.

To our knowledge, our work is the largest attempt to study the reliability and validity of these tools -only Downs and Black (1998) analyzed their tool's reliability and validity applying it to 10 prospective studies with worse results than ours-. However, our results should be considered with caution because of several reasons. First, while the tool by Downs and Black (1998) originally considers the use of a summary score, neither Fowkes and Fulton (1991) nor Berra *et al.* (2008) do. Instead, they suggest a subjective evaluation of the responses given to their items. In this study, and in order to be able to make comparisons, we decided to compute the global scores, which may have leaded to different results than if a subjective assessment was used.

Second, the maximum score for some domains was very low when using the tool by Downs and Black (1998) because of the low number of items covering these domains, the high number of not applicable items to certain research designs, and the mainly dichotomous response style of the tool. This leaded in some cases to a low or absent variability among scores, making the agreement coefficients prone to be low or incalculable.

Third, the clearly different patterns for the two raters observed in the agreement scores between tools raise some reflections. Indeed, since inter-rater agreement is in general low it is not strange that the agreement coefficients between tools do not match from one rater to another. What is confusing, though, is that for one rater all tools had moderate statistically significant agreement coefficients when comparing the global scores. The most evident difference among the two raters is their experience in methodology, since one of them is a graduate student in this field, while the second one is an associate professor. Since wide, double-barreled, and high-inference items were the rule rather than the exception (and instructions scarce and not always clarifying), rater one could have interpreted items as literally as possible, while rater two could have relied more on his background knowledge to make higher inferences. Anyway, although the influence of the different expertise between raters cannot be discarded, it is true that none of the applied quality assessment tools required that their users should have any specific knowledge in this field. So, if knowledge in methodology of the tools' users substantially affects their assessment of quality, concern rises about their usage across systematic reviews and meta-analyses. With that said, we acknowledge that we expected a higher agreement between tools, considering that they were chosen because they were the tools that had the widest coverage of our domains.

Finally, we have no clear explanation why all tools had such a good inter-rater agreement when applied to cross-sectional studies, especially considering the results in the other two designs. Although the number of not applicable items was higher when cross-sectional studies were assessed, we do not think that this difference could explain itself the good inter-rater agreement.

In conclusion, it is difficult to recommend without reservation a tool for assessing the methodological quality of studies that have either a prospective design or a retrospective design with quasi-control group. In this sense, although the tool by Downs and Black (1998) showed a moderate inter-rater reliability for the global score, this did not consistently happen at the domain's level. On the other hand, the tools by Dows and Black (1998) and Berra *et al.* (2008) stand out when the assessed studies have cross-sectional designs. Despite having wide, double-barreled and high-inference items, these two tools have a remarkable inter-rater reliability both for the global score and for most of the domains of quality. Moreover, the fact that the tool by Berra *et al.* (2008) is written in Spanish might limit its usability for non-Spanish speakers. Finally, although the tool by Fowkes and Fulton (1991) also has good inter-rater agreement scores for cross-sectional designs, we are reluctant to recommend it yet, as we consider that their behavior on the other designs demands more exhaustive testing.

Each tool had items related to all domains (except the domain Funding), which have let us infer 30 aspects that refine our domains of quality. These domains and aspects can be used as starting point to develop a new quality assessment tool of prospective, retrospective with quasi-control group, and cross-sectional studies following the established procedure that any assessment tool requires.

References

American Psychological Association (2010). *Publication Manual of the American Psychological Association, Sixth Edition*. Washington, D.C.: American Psychological Association.

Berra, S., Elorza-Ricart, J.M., Estrada, M.D., and Sánchez, E. (2008). A tool for the critical appraisal of epidemiological cross-sectional studies. *Gaceta Sanitaria*, *22*, 492-497.

Carretero-Dios, H. and 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*, 863–882.

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

Downs, S.H. and 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*, 377-384.

Fernández-Ríos, L. and Buela-Casal, G. (2009). Standards for the preparation and writing of Psychology review articles. *International Journal of Clinical and Health Psychology*, *9*, 329–344.

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

Jarde, A., Losilla, J.M., and Vives, J. (in press). Methodological quality assessment tools of non-experimental studies: A systematic review. *Anales de Psicología*.

Jüni, P., Altman, D.G., and 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., and Egger, M. (1999). The hazards of scoring the quality of clinical trials for meta-analysis. *Journal of the American Medical Association*, *282*, 1054-1060.

Kendall, M.G. (1938). A New Measure of Rank Correlation. *Biometrika*, *30*, 81-93.

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

Montero, I. and León, O.G. (2007). A guide for naming research studies in Psychology. *International Journal of Clinical and Health Psychology*, *7*, 847-862.

Sanderson, S., Tatt, I.D., and 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*, 666-676.

Shrier, I., Boivin, J.F., Steele, R.J., Platt, R.W., Furlan, A., Kakuma, R., Brophy, J., and Rossignol, M. (2007). Should meta-analyses of interventions include observational studies in addition to randomized controlled trials? A critical examination of underlying principles. *American Journal of Epidemiology*, *166*, 1203-1209.

Shrout, P.E. and Fleiss, J.L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420-428.

Valentine, J.C. and 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*, 130-149.

Vandenbroucke, J.P., von Elm, E., Altman, D.G., Gøtzsche, P.C., Mulrow, C.D., Pocock, S.J., Poole, C., Schlesselman, J.J., and Egger, M. (2007). Strengthening the reporting of observational studies in epidemiology (STROBE): Explanation and elaboration. *Epidemiology*, *18*, 805-835.

Wells, K. and Littell, J.H. (2009). Study quality assessment in systematic reviews of research on intervention effects. *Research on Social Work Practice*, *19*, 52-62.

West, S., King, V., Carey, T.S., Lohr, K.N., McKoy, N., Sutton, S.F., and Lux, L. (2002). Systems to rate the strength of scientific evidence. *Evidence Report/Technology Assessment*, 47, 1-11.

Received May 2, 2011
Accepted July 27, 2011

3.3 Q-Coh: A Tool to Screen the Methodological Quality of Cohort Studies in Systematic Reviews and Meta-analyses (Article 3)

Once the need for a valid and reliable quality assessment tool was established for at least cohort and case-control studies, we started developing a new tool following a rigorous procedure (which was the third of the specific objectives of this thesis). One of the first decisions we took was to focus only on one research design: cohort studies. We adapted the domains of quality we had used so far to the more well-known structure based on the main biases threatening a cohort study (selection bias, performance bias, detection bias, and attrition bias) and made a first draft of the structure of the tool based on the pool of items used by previous quality assessment tools (Annex 3). As the structure of the developing tool required the user to make judgments, we also wrote a manual with enough background information and instructions to help the users to take decisions with solid ground. This manual could not be published in a journal article (for obvious reasons), but has been attached as an annex in this thesis (Annex 4). The final version of the tool, which we named Q-Coh (Quality of Cohort studies), is an Excel spreadsheet, which allows a more user-friendly interface and summarizes the previously answered items (Annex 5), therefore avoiding that the user has to remember his previous answers or look them up when he has to make any judgments.

To assess its psychometric properties (which was the fourth specific objective of this thesis), we applied the Q-Coh to 21 studies of varying quality. Regarding validity, since there is no gold standard with which to assess the validity of our tool we studied the agreement of our ratings of the overall quality of the studies with an external rating (the classification of quality given by other authors using different assessment tools and/or procedures). Additionally we analyzed the degree of overlapping between our tool and the aspects covered in the bank of items based on all tools found in the systematic review of our first article.

The results showed that the proportion of agreement between pairs of raters was over 80% in all cases, with not only good to very good kappa values, but also being statistically significant in most inferences. This is very positive, especially considering the existing difficulties in developing a quality assessment tool with acceptable reliability scores.

Regarding the Q-Coh's validity, this result shows a moderate agreement between our tool and the classification of the studies made by other authors using different tools and procedures. In addition, almost all aspects appearing in the bank of items were covered by our tool.

The results of this study were presented at the V European Congress of Methodology in Santiago de Compostela (2012) and the paper was accepted for publication in the International Journal of Clinical and Health Psychology on December 2012. Its appendix has been attached as Annex 6.

Q-Coh: A tool to screen the methodological quality of cohort studies in systematic reviews and meta-analyses[1]

Alexander Jarde[2] (*Universitat Autònoma de Barcelona, Spain*), Josep-Maria Losilla (*Universitat Autònoma de Barcelona, Spain*), Jaume Vives (*Universitat Autònoma de Barcelona, Spain*), and Maria F. Rodrigo (*Universitat de València, Spain*)

ABSTRACT. The evaluation of the methodological quality of primary studies in a systematic review is a key process to enhance the likelihood of achieving valid results. When considering non-randomized designs as cohort studies, this process becomes even more critical, since these designs are more susceptible to bias than randomized controlled trials are. Taking this into account, a tool, named Q-Coh, was designed with the aim to screen the methodological quality of the primary studies with a cohort design priming specificity over sensitivity in a reasonable application time. After applying it to 21 prospective cohort studies by three raters, all domains had a moderate to good agreement, with all except one of them having statistically significant kappa values. Despite there is no gold standard for the methodological quality, arguments supporting its validity are given. Future research should assess the psychometric properties of Q-Coh in the context of real meta-analyses, evaluate the influence of the raters' substantive and methodological expertise on these properties, and explore different ways of including the domains-based ratings of the quality provided by Q-Coh into meta-analyses.

KEYWORDS. Quality. Cohort studies. Systematic review. Meta-analysis. Instrumental study.

RESUMEN. La valoración de la calidad metodológica de estudios primarios en una revisión sistemática es un proceso clave para mejorar la validez de los resultados. Al considerar diseños no aleatorizados como los estudios de cohortes, este proceso se vuelve aún más crítico, ya que estos diseños son más susceptibles a sesgos que los estudios controlados mediante aleatorización. Teniendo esto en cuenta se diseñó Q-Coh, una herramienta cuyo objetivo es valorar la calidad metodológica de estudios primarios con un diseño de cohortes, primando la especificidad sobre la sensibilidad y con un tiempo de aplicación razonable. Después de ser aplicada a 21 estudios de cohortes por tres evaluadores, todas las dimensiones obtuvieron un acuerdo entre moderado y bueno, teniendo todas excepto una de ellas valores de kappa estadísticamente significativos. A pesar de no existir ningún criterio de referencia estándar para valorar la calidad metodológica, se dan argumentos que respaldan la validez de Q-Coh.

[1] This research was supported by Grant PSI2010-16270 from the Spanish Ministry of Science and Innovation.

[2] Correspondence: 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

Investigaciones futuras deberán estudiar las propiedades psicométricas de la herramienta en el contexto de meta-análisis reales, evaluar la influencia de los conocimientos sustantivos y metodológicos de los evaluadores sobre dichas propiedades, y explorar diferentes vías para incluir en los meta-análisis las puntuaciones de calidad de las dimensiones proporcionados por Q-Coh.

PALABRAS CLAVE. Calidad. Estudios de cohortes. Revisión sistemática. Meta-análisis. Estudio instrumental.

The evaluation of the methodological quality of primary studies in a systematic review and meta-analyses is a key process to enhance the likelihood of achieving valid results. When considering non-randomized designs as cohort studies, this process becomes even more critical, since these designs are more susceptible to bias than randomized controlled trials are. Among the so-called "observational studies" in the epidemiological tradition (ex post facto studies in Montero and León's nomenclature; 2007), where the researcher does not carry out any intervention, cohort studies are always considered as having the highest internal validity. Dozens of tools have been developed up to date to assess the quality of prospective studies, but there's no clear candidate to be recommended without doubts. In fact, all systematic reviews collecting this type of tools (Deeks et al., 2003; Jarde, Losilla, & Vives, 2012a; Sanderson, Tatt, & Higgins, 2007; Shamliyan, Kane, & Dickinson, 2010; West et al., 2002) agree in criticizing that most of them have not been developed using standard psychometric techniques. This issue has been addressed in the last years and there have been initiatives to explore the psychometric properties of already existing tools (Jarde, Losilla, & Vives, 2012b) and new proposals of assessment tools of methodological quality have been developed using more rigorous procedures (*e.g.*, Shamliyan, Kane, Ansari, et al., 2010; Viswanathan & Berkman, 2012). Jarde et al. (2012b) applied three tools highlighted in a previous systematic review (Jarde et al., 2012a) to 30 studies with prospective, retrospective and cross-sectional designs, but found low inter-rater reliability in prospective studies. Similarly, Shamliyan, Kane, Ansari, et al. (2010) and Viswanthan and Berkman (2012) developed their tools using a structured procedure but had poor agreement between raters.

The objective of this study is to develop a valid and reliable tool to be used in systematic reviews and meta-analyses to screen the methodological quality of primary studies with cohort designs.

Method

Purpose of the tool and the scope of the construct to be measured

The purpose of this tool, which has been named "Q-Coh" (Quality of Cohort studies), is to identify those cohort studies with low quality and therefore potential source of bias in the meta-analysis. It is not meant to be exhaustive, since there are aspects of the study's quality which might be too complex and variable (depending on the topic under study) to be assessed precisely with a closed tool, as for example the assessment of statistical analyses. Therefore, the Q-Coh tool focuses on the more essential aspects to set an acceptable level of methodological quality that a study should have, priming specificity over sensitivity.

Several overlapping terms have been used to define the construct to be measured by assessment tools of methodological quality, including internal/external validity, risk of bias, study limitations, precision, etc. (Viswanathan & Berkman, 2012). However, with the appearance of communication guidelines as the STROBE Statement (Vandenbroucke et al., 2007), it has been increasingly clear that an assessment tool of methodological quality should not address quality of reporting. Instead, it is argued that these tools have to focus on internal validity (Dreier, Borutta, Stahmeyer, Krauth, & Walter, 2010). What is less clear, though, is if external validity should or should not be assessed.

In this study, the construct labeled as *methodological quality* (or just *quality*), refers to the degree to which the study employs procedures to guarantee that the comparability of the groups is maintained along the whole study (and/or controlled for in the analyses), that the measures and results are valid and reliable, and that the results can be extrapolated to the target population. Therefore, this construct does not include aspects related to the correctness or completeness of the studies' reporting, nor is related to other aspects considered of good research practice, but that are not susceptible to introduce systematic differences between the groups compared in the studies (*e.g.* ethical committee's approval, sample size/power calculation).

Regarding the definition of cohort studies, in the STROBE statement cohort studies are described as follows:

> In cohort studies, the investigators follow people over time. They obtain information about people and their exposures at baseline, let time pass, and then assess the occurrence of outcomes. Investigators commonly make contrasts between individuals who are exposed and not exposed or among groups of individuals with different categories of exposure. Investigators may assess several different outcomes, and examine exposure and outcome variables at multiple points during follow-up. (Vandenbroucke et al., 2007)

Therefore, cohort studies as described by the STROBE statement would be classified as types of ex post facto studies in Montero and León's (2007) classification of research studies. In fact, the definition of cohort studies is not straightforward, since authors and databases of different fields use a variety of terminology. So, 'longitudinal study', 'follow-up study', 'cohort study' and 'prospective study' are closely related terms and are commonly used as synonyms (Vandenbroucke et al., 2007). This might not be surprising considering that the definitions and relations between them are not consistent along different reference sources. Given this heterogeneity, the Cochrane Non-Randomized Studies Methods Group advises those authors interested in including non-randomized studies in their reviews not to rely on design labels, but to use explicit study design features (Higgins & Green, 2011). Therefore, in this work any study with the following characteristics will be considered a cohort study: 1. There is a comparison between at least two groups to assess the effect of an exposure on an outcome. 2. The groups are defined by the exposure variable. 3. On onset, none of the participants has the outcome of interest. 4. Investigators do not handle who is exposed or not. 5. Information about the exposure and the outcome is not registered concurrently. There may be studies that do not satisfy these characteristics but that are considered as 'prospective studies' by other authors. It is not this paper's intention to open a discussion about that. The work presented here will simply not be appropriate for those studies.

Tool's specifications

The study focuses only on cohort studies, because, on one hand, they have some design characteristics not shared with retrospective and cross-sectional studies, therefore avoiding an omnibus tool. On the other hand, the wide array of topics and areas where cohort designs are applied makes the task challenging enough, especially considering the difficulties found in previous initiatives to obtain good reliability scores.

Forcing the same response options pattern to all items was avoided, since not all response options are always suited for all items. For example, a common response option often appearing by default in other assessment tools is the 'Not Reported' option, which can be very confusing or unnecessary in certain cases. Therefore, although an effort was made to maintain the response options homogeneous, each item was given the response options that fitted the potential answers best. Additionally, the response options were polarized as much as possible, avoiding gradients (*e.g.* yes, somehow, no), avoiding an ambiguous 'comfort zone' response, and forcing the user to make either a positive or a negative judgment in the inferences.

Development and testing of the Q-Coh

A bank of items was built with all the items of the tools located in a previous systematic review of assessment tools of methodological quality for non-randomized studies (Jarde et al., 2012a). The items were grouped into seven domains which assess representativeness, comparability of the groups at the beginning of the study, quality of the exposure measure, maintenance of the comparability during the follow-up time, quality of the outcome measure, attrition, and statistical analyses. These domains were derived from the extended classification of biases (selection bias, performance bias, detection bias, and attrition bias). Finally, those items asking for details not required by either the STROBE statement (Vandenbroucke et al., 2007) or the Journal Article Reporting Standards (American Psychological Association, 2010) were discarded. This process resulted in a first draft with 55 items and 7 inferences; and a response manual with instructions and additional information to answer the items.

This draft was revised and reduced to a pilot version of the tool with only 29 items and 7 inferences by combining some highly atomized items or straightforward inferences, making some higher inferences and deleting some items considered too specific (mostly regarding the statistical analyses). Additionally, the user manual was integrated into the Q-Coh, indicating when to answer which response option and making clarifying comments when needed. Finally, five items were included at the beginning of the tool to check for the characteristics that define a cohort study to assess if the tool is applicable or not in each case.

In order to have a list of studies to apply the Q-Coh tool to, a pool of cohort studies was made with studies that were previously used in published meta-analyses and whose quality had somehow been assessed. Therefore, each study was classified into low, acceptable or good quality based on the evaluation it had received by the reviewers. The order in which the studies were evaluated was at random and the reviewers were blinded to their classification of quality.

After the pilot version of the Q-Coh was applied to three studies (one of each level of quality) by three of the authors (AJ, JV, MFR), all specialists in the field of research methodology, a final version of the tool with 26 items and 7 inferences (plus five initial items to check for the characteristics of the study design) was developed (see Table 1). The same three authors applied this final version to 21 articles (7 of each level of quality). These articles were from different topics, including obesity, depression, childhood abuse, Alzheimer disease, job satisfaction, and menopausal transition among others. To deal with this heterogeneity, a common target population (inference 1) was defined, as well as the level of precision required for considering the selection criteria ‘explicit’ (item 4), and when to consider a confounding factor ‘important’ (items 7 and 13). For the same reason, the assessment of the overall quality was made using the following algorithm: When none or one domain were evaluated negatively, the overall quality was considered good. If two domains were evaluated negatively, the overall quality was considered acceptable. Finally, if more than two domains were evaluated negatively, the overall quality was considered low.

TABLE 1. Domains, Items and Inferences of the Q-Coh (with response options).

Design of the study

Item.A. Is there a comparison between at least two groups to assess the effect/ association of an exposure and an outcome? (Yes/No)

Item.B. Are the groups defined by the exposure variable? (Yes/No)

Item.C. Has or could any of the participants have the outcome of interest on onset? (No/Yes)

Item.D. Do investigators handle who is exposed or not? (No/Yes)

Item.E. Is information about the exposure and the outcome of interest registered concurrently? (No/Yes)

Inference.0. Is the tool suitable for this study? (Yes/No)

Representativeness

Item.1. Have the study participants been selected using a randomized sampling procedure? (Yes/No)

Item.2. Is the similarity between the selected group of subjects and the target population justified by the authors? (Yes, empirically/Yes, verbally/No)

Item.3. Is there a predominant reason for refusing to participate at the beginning of the study? (No-Irrelevant/Yes/Not reported)

Inference.1. Could the results be generalized from the sample to the target population? (Probably/Unlikely)

Comparability of the groups

Item.4. Were the inclusion and exclusion criteria explicitly defined for all groups? (Yes/No)

Item.5. Were the same inclusion and/or exclusion criteria applied equally to all groups? (Yes/No/Not Reported)

Item.6. Could differences in the selection criteria introduce systematic differences between the groups (other than exposure)? (Unlikely/Probably)

Item.7. Were known confounding factors accounted for in the design or in the analysis? (Yes/Partially/No)

Inference.2. Is bias between the groups avoided at the beginning of the study? (Probably/Unlikely)

Exposure measure

Item.8. Was the exposure explicitly defined? (Yes/No)

Item.9. Was the tool used to measure the exposure variable valid? (Yes/Presumably/Doubtfully)

Item.10. Was the tool used to measure the exposure variable reliable? (Yes/Presumably/Doubtfully)

Item.11. Was the procedure to measure the exposure the same for all participants? (Yes/No/Not Reported)

Inference.3. Could the classification of the participants into exposed or unexposed be biased? (Unlikely/Probably)

Maintenance of the comparability

Item.12. Were potential confounders that appeared during the follow-up time taken into account in the analyses? (Yes/No)

Item.13. Was the length of follow-up similar between the groups? (Yes/No, but controlled/No)

Item.14. Is there any potential confounder that could have appeared during follow-up that was not taken into account by the authors? (Probably none important/Probably/Yes)

Inference.4. Could the exposure to other factors appearing during follow-up introduce systematic differences between the groups? (Unlikely/Probably)

Outcome measure

Item.15. Was the outcome variable explicitly defined? (Yes/No)

Item.16. Was the tool used to assess the outcome variable valid? (Yes/Presumably/No)

Item.17. Was the tool used to assess the outcome variable reliable? (Yes/Presumably/No)

Item.18. Was the tool used to assess the outcome appropriate? (Probably/Unlikely)

Item.19. Was the outcome variable assessed in the same way in all groups? (Yes/No)

Item.20. Was the outcome variable assessed at the same time for all groups? (Yes/No)

Item.21. Was the outcome variable assessed in the same context for all groups? (Yes/No)

Item.22. Could the procedures for measuring the outcome variable introduce systematic differences between the groups? (Unlikely/Probably)

Item.23. Were the participants successfully blinded to the research question? (Yes/No/Not necessary)

Item.24. Were those assessing the outcome successfully blinded to the exposure status of the participants? (Yes/No/Not necessary)

Inference.5. Does the measure of the outcome variable reflect the true situation? (Probably/Unlikely)

Attrition

Item.25. Were drop out rates similar in all groups? (Yes/No/Not Reported)

Item.26. Were reasons for dropping out similar in all groups? (Yes/No/Not Reported)

Inference.6. Could incomplete information introduce systematic differences between groups? (Unlikely/Probably)

Statistical analyses

Inference.7. Do the results of the statistical analysis reflect the true situation? (Probably/Unlikely)

Overall assessment of the study’s quality

What overall quality does this study have? (Good /Acceptable /Low)

Note. The original tool is a spreadsheet that allows recording the responses, has the instructions embedded, and reminds the answers made to the previous items that have to be considered in some cases. This spreadsheet version of the Q-Coh can be requested to the authors.

Since there is no gold standard with which to assess the validity of the tool only an approximation is possible. Therefore, the validity of the Q-Coh was analyzed by studying the agreement of the ratings of the overall quality of the studies with an external rating: the classification of quality given by other authors using different assessment tools and/or procedures.

On the other hand, the bank of items reflects all aspects considered previously in the assessment of the quality of cohort studies. Considering that lots of these items have been developed by methodological experts, it is very unlikely that there is any important aspect that is not considered in the bank of items developed for this study. Therefore, analyzing the degree of overlapping between the Q-Coh and the aspects covered in the initial bank of items shall give an idea of the validity of the tool. Of the 57 aspects covered by the bank of items, 39 were considered by the Q-Coh tool and 18 were not. The reasons why these aspects were not covered in the tool were because they assessed aspects not related to the definition of quality proposed here (three aspects regarding reporting, one aspect regarding sample size), because they were too specific (three aspects not considered by the STROBE statement, four aspects assessing details of the statistical analyses) or too broad (one aspect referring to quality control procedures in general). Therefore, six aspects (11%) of the bank of items that were not covered by our tool remain open for discussion: Funding, conflicts of interests, memory biases, contamination, follow-up time, and appropriateness of the evaluation methods.

Statistical analyses

In order to evaluate the inter-rater agreement between two raters the Cohen's kappa coefficient (Cohen, 1960) or its generalization for multiple raters as the one proposed by Fleiss (1971) traditionally have been the most widely used statistics. However, these statistics are not recommended when the prevalence of a given response category is very high or low. In this situation the "kappa paradox" (Feinstein & Cicchetti, 1990) takes place so that the value of the kappa statistic is low even when the observed proportion of agreement is quite high. A second kappa paradox results from the influence of bias in the kappa value. Bias refers to the extent to which the raters disagree on the proportion of cases in each response category. When there is a large bias, kappa is higher than when bias is low or absent. Given that kappa is difficult to interpret in presence of different prevalence or bias, several studies have recommended reporting other statistics, in addition to kappa, to describe more thoroughly the extent of agreement between raters and the possible causes of disagreement. For instance, some authors have recommended informing about the proportions of specific agreement between raters for each response category to evaluate the possible effect of prevalence or bias (Cicchetti & Feinstein, 1990; Lantz & Nebenzahl, 1996; Uebersax, 2012). Additionally, in presence of different prevalence or bias a widely used alternative to Cohen's kappa is the Prevalence-Adjusted and Bias-Adjusted Kappa (PABAK) proposed by Byrt, Bishop, and Carlin (1993).

In this paper several statistics are given for each item. The proportion of agreement between the three raters, the proportion of agreement between pairs of raters, the proportion of choices of the three raters and the proportion of agreement between pairs of raters for each response category; and the Fleiss kappa statistic (or PABAK when necessary). All these analyses have been performed using the "irr" package (v.0.83) for R version 2.15.0 (Gamer, Lemon, Fellows, & Singh, 2012).

As already mentioned, to assess the validity of the Q-Coh, the agreement between the three rater's assessment of the studies' global quality and the external rating of quality based on the assessment made by the authors of the meta-analyses where the studies were located was analyzed. In addition, the association between these external ratings and the number of domains evaluated negatively by the three raters for each study was also evaluated. In both cases, to obtain a unique rating the majority criterion was applied. These analyses were performed using the Weighted Cohen's Kappa (Fleiss, Cohen, & Everitt, 1969) and the nonparametric Kendall tau-b (τb) correlation coefficient (Kendall, 1938), respectively.

Results

Inter-rater reliability

Following Landis and Koch's (1977) criteria, the agreement was good to very good in all inferences evaluating the different domains of quality (kappa: .68 to .87) except for *Attrition* (kappa = .60); with a proportion of agreement between pairs of raters ranging from 81% to 94% (71% to 90% between all three raters); and similar rates of agreement were found at the items level. The overall assessment of quality was good (kappa = .75), with a proportion of agreement between pairs of raters of 87% (86% between all three raters). All kappa values of the domains were statistically significant except for the inference assessing the domain *Outcome measure*. Table 2 summarizes the results of the agreement analyses.

On other hand, in four items of the domain *Outcome measure* the kappa was not applicable due to a lack of variability, since all raters answered the same response category in all cases. Similarly, over 90% of the responses were concentrated in one single response category in two items belonging to the domain *Exposure measure*, two items belonging to *Outcome measure*, and in one item of the domain *Attrition*. The inference *Outcome measure* shows also a remarkable lack of variability as 97% of the responses are concentrated in one category.

Finally, the domain *Statistical analyses* consists of a single inference. It has a very good and statistically significant value of kappa (.87), but there is also very little variability in the answers given.

Validity

To evaluate the agreement between the three rater's assessment of the studies' global quality and the external ratings of quality of these studies, a weighted kappa was applied, with weights [0, 1, 3], that resulted in a value equal to 0.41 (p = .035). This result shows a moderate agreement between both ratings. Moreover, to evaluate the association between the external ratings of quality and the number of dimensions evaluated negatively by the three raters for each study, we compute the Kendall tau-b (τb) correlation coefficient, which results in a value equal to -0.454 (p = .003). This value indicates an inverse association between both variables, *i.e.*, a high number of domains negatively evaluated is associated with a low global quality rating.

Discussion

The proportion of agreement between pairs of raters is over 80% in all cases, with not only good to very good kappa values, but also being statistically significant in most inferences. Considering the existing difficulties in developing a reliable tool for assessing the methodological quality of non-randomized studies in general, these are very positive results. Another strength, besides its psychometric properties, is the fact that the Q-Coh checks for its applicability to the considered study by assessing its design characteristics in the initial items. Additionally, the reduced number of items and the instructions embedded into the tool make this tool feasible to apply even in large reviews with users with low methodological expertise in a reasonable amount of time.

While there are domains that can be assessed without a defined context, it is necessary that certain criteria are established a priori to assess some of the domains, as suggested by other authors (*e.g.* Shamliyan, Kane, Ansari et al., 2010; Valentine & Cooper, 2008). Therefore, to assess the comparability of the groups and its maintenance along the follow-up period, the list of confounders considered important to be controlled has to be defined. Additionally, the criteria that should be used to make the overall assessment of the quality (whether it is appraised as another inference or by applying an algorithm) should be discussed before applying it, too.

TABLE 2. Results of agreement analyses.

Domains' items and inferences	P. Total agreement (3 raters)	P. Overall agreement (2 raters)	Response category 1			Response category 2			Response category 3			Kappa
			Response	P. cat.	P. agree.	Response	P. cat.	P. agree.	Response	P. cat.	P. agree.	
Study Design												
Item.A	1.0	1.0	*Yes*	1.0	1.0	*No*	.00					
Item.B	.95	.97	*Yes*	.98	.98	*No*	.02	.00				.94^{a}
Item.C	.90	.94	*No*	.97	.97	*Yes*	.03	.00				.87^{a}
Item.D	1.0	1.0	*No*	1.0	1.0	*Yes*	.00					
Item.E	.95	.97	*No*	.98	.98	*Yes*	.02	.00				.94^{a}
Inference.0	.86	.90	*Yes*	.95	.95	*No*	.05	.00				.81^{a}
Representativeness												
Item.1	.90	.94	*Yes*	.10	.67	*No*	.90	.96				.87^{a**}
Item.2	.76	.84	*Yes, empirical*	.02	.00	*Yes, verbally*	.17	.55	*No*	.81	.92	.68^{a**}
Item.3	.76	.84	*No/Irrelevant*	.14	.44	*Yes*	.00		*NR*	.86	.91	.68^{a*}
Inference.1	.81	.87	*Probably*	.29	.78	*Unlikely*	.71	.91				.75^{a**}
Comparability of the groups												
Item.4	.71	.81	*Yes*	.67	.86	*No*	.33	.71				.62^{a**}
Item.5	.86	.90	*Yes*	.67	.93	*No*	.03	.00	*NR*	.30	.95	.79**
Item.6	.81	.87	*Unlikely*	.87	.93	*Probably*	.13	.50				.75^{a**}
Item.7	.71	.81	*Yes*	.63	.85	*Partially*	.37	.74	*No*	.00		.59**
Inference.2	.76	.84	*Probably*	.79	.90	*Unlikely*	.21	.62				.68 ^{a**}
Exposure measure												
Item.8	.95	.97	*Yes*	.97	.98	*No*	.03	.50				.94^{a**}
Item.9	.76	.84	*Yes*	.52	.91	*Presumably*	.38	.79	*Doubtfully*	.10	.67	.72**
Item.10	.71	.81	*Yes*	.56	.89	*Presumably*	.30	.68	*Doubtfully*	.14	.78	.67**
Item.11	.95	.97	*Yes*	.98	.98	*No*	.00		*NR*	.02	.00	.94^{a}
Inference.3	.81	.87	*Unlikely*	.87	.93	*Probably*	.13	.50				.75 ^{a**}
Maintenance of comparability												
Item.12	.71	.81	*Yes*	.27	.65	*No*	.73	.87				.62^{a**}
Item.13	.90	.94	*Yes*	.89	.96	*Controlled*	.11	.71	*No*	.00		.87^{a**}
Item.14	.81	.87	*P. none imp.*	.44	.86	*Probably*	.56	.89	*Yes*	.00		.74*
Inference.4	.81	.87	*Unlikely*	.48	.87	*Probably*	.52	.88				.75**

Domains' items and inferences	P. Total agreement (3 raters)	P. Overall agreement (2 raters)	Response category 1			Response category 2			Response category 3			Kappa
			Response	P. cat.	P. agree.	Response	P. cat.	P. agree.	Response	P. cat.	P. agree.	
Outcome measure												
Item.15	1.0	1.0	*Yes*	1.0	1.0	*No*	.00					
Item.16	.71	.81	*Yes*	.54	.82	*Presumably*	.46	.79	*Doubtfully*	.00		.62**
Item.17	.71	.81	*Yes*	.52	.82	*Presumably*	.48	.80	*Doubtfully*	.00		.62**
Item.18	1.0	1.0	*Probably*	1.0	1.0	*Unlikely*	.00					
Item.19	1.0	1.0	*Yes*	1.0	1.0	*No*	.00					
Item.20	.81	.87	*Yes*	.94	.93	*No*	.06	.00				.75[a]
Item.21	1.0	1.0	*Yes*	1.0	1.0	*No*	.00					
Item.22	.95	.97	*Unlikely*	.98	.98	*Probably*	.02	.00				.94[a]
Item.23	.76	.84	*Yes*	.14	1.0	*No*	.13	.38	*Not necess.*	.73	.89	.68[a]**
Item.24	.76	.83	*Yes*	.30	.95	*No*	.16	.50	*Not necess.*	.54	.85	.71**
Inference.5	.90	.94	*Probably*	.97	.97	*Unlikely*	.03	.00				.87[a]
Attrition												
Item.25	.71	.81	*Yes*	.19	.50	*No*	.05	1.0	*NR*	.76	.88	.62[a]**
Item.26	.90	.94	*Yes*	.03	.00	*No*	.00		*NR*	.97	.97	.87[a]
Inference.6	.71	.81	*Unlikely*	.60	.84	*Probably*	.40	.76				.60*
Statistical analyses												
Inference.7	.90	.94	*Probably*	.95	.97	*Unlikely*	.05	.33				.87[a]*
Overall assessment of quality												
Overall	.86	.87	*Good qual.*	.17	.82	*Acceptable*	.05	.00	*Low qual.*	.78	.94	.75[a]**

Note. P. Total agreement (3 raters) = proportion of agreement between all of the three raters; P. Overall agreement (2 raters) = proportion of agreement between pairs of raters; P. cat. = proportion of choices of the three raters for a specific response category; P. agree. = proportion of specific agreement between pairs of raters for each response category; Kappa = Fleiss Kappa (or PABAK).

[a] PABAK.

* $p < .05$. ** $p < .01$

The Q-Coh was applied to a relatively high number of studies (compared to other validations of similar tools), making sure a wide spectrum of study quality was covered. This resulted in a wide array of topics addressed by these studies. The fact that despite this variety in addressed topics the agreement between raters was generally good suggests that the tool is flexible enough to be applied across topics maintaining an acceptable inter-rater reliability. This is probably so because the tool requires to make an a priori definition of the topic-dependent criteria.

Some items were not very discriminant, since all or most of the answers were the same for all studies. In some items (8, 11, 15, 18 to 22) the predominant response reflects a positive value. However, the fact that most of the studies score positively does not mean that these aspects could be left out, since a negative assessment could severely downgrade the study's methodological quality. In item 26 the predominant response was 'Not Reported'. This item deals with the reasons given for abandoning the study. The fact that this information is not reported is probably not because of a bad reporting in most cases, but because that information is not available to the researchers. Q-Coh could also be used in this sense to check the reporting quality of the manuscripts prior to their publication.

Regarding the aspects of the bank of items not covered by the tool, the most notable are probably the ones referring to funding and conflicts of interest. Despite it is a common critique made to tools of this kind, these aspects were excluded because it was considered that they do not require any additional item. Indeed, although funding and conflicts of interest can influence the quality of the study at many of its stages, the tool already assesses each stage separately in its domains. Moreover, the funding is not the only source of conflicts of interest, as personal, academic or political interests, which are rarely reported, could also be affecting the quality of a study.

In order to improve the Q-Coh tool, future studies should focus, on one hand, on enhancing the inter-rater reliability. On the other hand, the tool's psychometric properties should be assessed in the context of real systematic reviews and meta-analyses, and with other raters with substantive and methodological expertise.

Finally, going beyond the screening use of the Q-Coh, it would be interesting to explore the inclusion of the domains-based ratings of the quality provided by the Q-Coh into meta-analyses. How exactly this should be done is still under discussion. Detsky, Naylor, O'Rourke, McGeer, and L'Abbé. (1992) have suggested four ways of doing so when the methodological quality has been summarized in a single overall quality score, and Thompson et al. (2010) have proposed to include into meta-analyses a quantification of the extent of internal and external biases. All these suggestions may be a good starting point to work in.

References

American Psychological Association (2010). *Publication Manual of the American Psychological Association (6th edition).* Washington, D.C.: American Psychological Association (APA).

Byrt, T., Bishop, J., & Carlin, J.B. (1993). Bias, prevalence and Kappa. *Journal of Clinical Epidemiology*, *46*, 423-429.

Cicchetti, D.V. & Feinstein, A.R. (1990). High agreement but low Kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, *6*, 551–558.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37-46.

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

Detsky, A.S., Naylor, C.D., O'Rourke, K., McGeer, A.J., & L'Abbé, K.A. (1992). Incorporating variations in the quality of individual randomized trials into meta-analysis. *Journal of Clinical Epidemiology, 45*, 255-265.

Dreier, M., Borutta, B., Stahmeyer, J., Krauth, C., & Walter, U. (2010). *Vergleich von Bewertungsinstrumenten für dis Studienqualität von Primär- und Sekundärstudien zur Verwendung für HTA-Berichte im deutschsprachigen Raum* [Comparison of tools for assessing the methodological quality of primary and secondary studies in health technology assessment reports in Germany] (HTA Bericht No. 102). Köln, Germany: Deutsche Agentur für Health Technology Assessment.

Feinstein, A.R. & Cicchetti, D.V. (1990). High agreement but low Kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, *6*, 543-549.

Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*, 378-382.

Fleiss, J.L., Cohen, J., & Everitt, B.S. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin, 72*, 323-327.

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2012). *Various Coefficients of Interrater Reliability and Agreement. Package «irr» for R* [Computer software]. Author.

Higgins, J.P. & Green, S. (2011). *Cochrane Handbook for Systematic Reviews of Interventions* (5.1.0 ed.). The Cochrane Collaboration. Available from www.cochrane-handbook.org.

Jarde, A., Losilla, J.M., & Vives, J. (2012a). Methodological quality assessment tools of non-experimental studies: a systematic review. *Anales de Psicología, 28*, 617-628.

Jarde, A., Losilla, J.M., & Vives, J. (2012b). Suitability of three different tools for the assessment of methodological quality in ex post facto studies. *International Journal of Clinical and Health Psychology*, *12*, 97-108.

Kendall, M.G. (1938). A New Measure of Rank Correlation. *Biometrika*, *30*, 8193.

Landis, J.R. & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159-174.

Lantz, C.A. & Nebenzahl, E. (1996). Behavior and interpretation of the κ statistic: resolution of the two paradoxes. *Journal of Clinical Epidemiology, 49*, 431-434.

Montero, I. & León, O.G. (2007). A guide for naming research studies in Psychology. *International Journal of Clinical and Health Psychology, 7*, 847-862.

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*, 666-676.

Shamliyan, T.A., Kane, R.L., Ansari, M.T., Raman, G., Berkman, N.D., Grant, M., et al. (2010). *Development quality criteria to evaluate nontherapeutic studies of incidence, prevalence, or risk factors of chronic diseases: pilot study of new checklists* (AHRQ Publication No. 11-EHC008-EF). Rockville, MD: Agency for Healthcare Research and Quality.

Shamliyan, T.A., Kane, R.L., & Dickinson, S. (2010). A systematic review of tools used to assess the quality of observational studies that examine incidence or prevalence and risk factors for diseases. *Journal of Clinical Epidemiology, 63*, 1061-1070.

Thompson, S., Ekelund, U., Jebb, S., Lindroos, A.K., Mander, A., Sharp, S., Turner, R., et al. (2010). A proposed method of bias adjustment for meta-analyses of published observational studies. *International Journal of Epidemiology, 40*, 765-777.

Uebersax, J. (2010). *Statistical Methods for Rater and Diagnostic Agreement: Recommended Methods*. Retrieved July 3rd 2012, from http://www.john-uebersax.com/stat/agree.htm.

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*, 130-149.

Vandenbroucke, J.P., von Elm, E., Altman, D.G., Gøtzsche, P.C., Mulrow, C.D., Pocock, S.J., et al. (2007). Strengthening the reporting of observational studies in epidemiology (STROBE): explanation and elaboration. *Epidemiology*, *18*, 805-835.

Viswanathan, M. & Berkman, N.D. (2012). Development of the RTI item bank on risk of bias and precision of observational studies. *Journal of Clinical Epidemiology*, *65*, 163-178.

West, S., King, V., Carey, T.S., Lohr, K.N., McKoy, N., Sutton, S.F., et al. (2002). *Systems to rate the strength of scientific evidence* (AHRQ Publication No. 02-E016). Rockville, MD: Agency for Healthcare Research and Quality.

Received July 26, 2012
Accepted December 5, 2012

4 Discussion

Along the different studies made in the frame of this thesis we have shown that, despite there are dozens of methodological quality assessment tools theoretically applicable to observational studies, there is none that satisfactorily covers all domains of quality that we consider important and that has good psychometric properties. Focusing on the quality of cohort studies, we have developed a quality assessment tool following the rigorous procedures that are standard for the development of any measurement instrument, including a careful analysis of its psychometric properties. This makes our tool an important candidate to be taken into account when systematic reviews and meta-analyses are performed including cohort studies, and fills an important gap in the research community.

At this point, it is interesting to run our checklist of the test specifications on our tool, the Q-Coh, in order to discuss in how far we have achieved our goals.

1. ***Rigorous development process.*** The Q-Coh was effectively developed following the established procedures. The construct to be measured was clearly defined, the generation of items followed an objective and systematic procedure, a pilot version of the tool was tested, and the reliability and validity of the final version has been assessed.

2. ***High inter-rater reliability.*** The proportion of agreement between pairs of raters is over 80% in all cases, with not only good to very good kappa values, but also being statistically significant in most inferences. The agreement between three raters was moderate to good in all domains, with all except one of them having a statistically significant kappa values. This is a fairly good inter-rater reliability, especially considering the reliability scores achieved with other tools in the literature.

3. ***No quantitative summary score (checklist, not a scale).*** Due to the heterogeneity of topics dealt by the articles assessed in our third study, the assessment of the overall quality was made using an algorithm instead of each rater making a judgment. However, the tool has no quantitative summary score.

4. ***Objective items.*** The first draft of the Q-Coh had extremely objective items requiring the presence of very concrete facts upon which several levels of inferences were build up. However, the pilot testing of this draft (a diagram of which has been attached as Annex 3) showed that this made the tool too long, complex and confusing. Therefore, some less precise terms were accepted (e.g. "Was the tool used to measure the exposure variable reliable?", or "Was the length of follow-up similar between the groups?"). However, the fact that the inter-rater reliability was good to very good in almost all items indicates that the items were objective enough.

5. ***Items ask for the presence of concrete facts.*** The Q-Coh has two types of items: those asking for the presence of concrete facts that can be retrieved from the assessed article directly; and inferences, which require the user to make a judgment based on the information extracted by the concrete items.

6. ***Instructions.*** The Q-Coh has detailed instructions embedded in it. It does not only give general instructions, but it indicates for each response option when it should be selected. In addition, the items that require the user to make a judgment (inferences) have also instructions guiding this decision-making process. Furthermore, in the manual a more extent information is provided, giving more background to each item and helping users with less methodological background to make their judgments.

7. ***Applicable by non-methodological experts.*** Although we made the manual and included clear instructions into the Q-Coh, our tool has not been applied by any non-methodological expert yet, so this point remains unchecked.

8. ***Moderate application time.*** The application time was about half an hour, which we consider appropriate.

9. ***Discern quality of reporting from study quality.*** As the reporting quality was explicitly excluded from our definition of quality, it was not assessed. Furthermore, the meaning of a not reported information was cautiously considered for each item. So, in some items a not reported information was considered a negative answer (e.g. Was the exposure explicitly defined?), while in others a 'Not reported' response option was necessary (e.g. Were the same inclusion and/or exclusion criteria applied equally to all groups?).

10. ***If external validity is assessed, it should be done separately from internal validity.*** This was done by considering a whole domain (Representativeness). As all the items (Items 1 to 3, plus Inference 1)

related to external validity are in this domain, it can be easily separated from the rest.

11. ***Hypothesis level taken into account.*** This is done by instructing the users to assess each outcome independently if more than one outcome is studied in the systematic review and/or the meta-analysis.

12. ***Generic enough to be applicable to different study fields.*** The Q-Coh has two types of items: those asking for the presence of concrete facts that can be retrieved from the assessed article directly; and inferences, which require the user to make a judgment based on the information extracted by the concrete items. It is this amount of subjectivity that is introduced into the evaluation process which gives the Q-Coh enough flexibility to be applied to different study fields. This is supported by the fact that the inter-rater agreement was good despite the variety of topics addressed by the studies to which it was applied.

13. ***Quality based on domains.*** The overall quality of a study is based in the Q-Coh on seven domains: assess representativeness, comparability of the groups at the beginning of the study, quality of the exposure measure, maintenance of the comparability during the follow-up time, quality of the outcome measure, attrition, and statistical analyses.

14. ***'Mixed-criteria' approach.*** The Q-Coh was designed with a 'mixed-criteria' approach, requiring objective facts regarding the study's design (the objective items), followed by a quality judgment (the inferences).

15. ***Standardized evaluation procedure.*** The specific instructions guide all the evaluation procedure. Even in the inferences, which require a more subjective assessment, it is specified what information should be considered when making the judgments (the previous objective items).

16. ***Transparent evaluation process.*** Because each inference has to be made based on the specified items it is easy to trace down the process by which a certain evaluation has been made.

17. ***Assessment of the magnitude and direction of bias.*** This has not been considered in the Q-Coh yet. This is mainly because our definition of quality focuses on the absence of bias, making an assessment of the magnitude and direction of bias not straightforward. In addition, we consider that the magnitude and direction of bias is more related to the inclusion of the tool's results into a meta-analyses, which is a topic that we still have not

addressed. Finally, we consider that making a valid and reliable tool to assess quality (absence of bias) was challenging enough. The assessment of the magnitude and direction of bias is a step further.

18. ***User-friendly interface.*** Although we consider that web-based interface or an Access form would have several benefits, we finally decided to make an Excel spreadsheet, since it can be shared more easily among researchers (it's a single file), and it's a file type everyone is familiar with and trusts (we deliberately avoided using macros to ensure this). A user-friendly interface was especially important in the Q-Coh, since it is not a list of items that just have to be answered in a row. On the contrary, there are items (inferences) which have to be answered based on the responses given on previous items. In order to avoid that the users have to look back, the interface presents them with a summary of the required items when they have to make an inference. Additionally, a system of colored icons also make the task of remembering the previously given responses easier.

Table 2. Summary checklist applied to the Q-Coh.

1	☑	Rigorous development process: - Definition of construct to be measured - Item generation - Pretest - Reliability and validity
2	☑	High inter-rater reliability.
3	☑	No quantitative summary score (checklist, not a scale)
4	☑	Objective items. Avoid imprecise terms.
5	☑	Items ask for the presence of concrete facts.
6	☑	Instructions: - Define the used terminology - Complete explanation of each item - Guidance when making decisions
7	▣	Objective enough to be applied by non-methodological experts.
8	☑	Moderate application time.
9	☑	Discern quality of reporting from study quality.
10	☑	If external validity is assessed, it should be done separately from internal validity.
11	☑	Hypothesis level taken into account.
12	☑	Generic enough to be applicable to different study fields.
13	☑	Quality based on domains.
14	☑	'Mixed-criteria' approach.
15	☑	Standardized evaluation procedure.
16	☑	Transparent evaluation process.
17	☒	Assessment of the magnitude and direction of bias.
18	☑	User-friendly interface.

5
Conclusion

The line of research started with this doctoral thesis is far from being closed and lots of new lines can, and hopefully will, stem from it. In the first place, although the results of the third study regarding the psychometric properties are promising, the Q-Coh can still be fine-tuned (e.g. including both the likely direction and magnitude of bias, or expanding the scope of applicable studies to those analyzing risk or prognostic factors) and its performance has to be assessed in other situations, as with other raters not involved in its development and raters with different methodological expertise. A second line of research that stems from this thesis is the exploration of different ways of including the results of the Q-Coh into the meta-analyses, for example making cumulative meta-analyses, using the domains' evaluations as moderate variables, etc. And a third line of research is of course the development of a tool to assess the methodological quality of case-control studies or a common tool applicable to both cohort and case-control studies, or even RCTs and quasi-RCTs.

The Q-Coh tool has been developed to be used in systematic reviews and meta-analyses, but it can also have other interesting uses. For example, journal editors could use it to ensure that only high quality cohort studies are published in their journal or that enough information is reported in order to make the assessment of its quality feasible. Another example of other uses could be as a support to teach students in psychology and health sciences about risk of biases, critical reading, etc.

6 Acknowledgements

I would like to thank Prof. Josep Maria Losilla for his invaluable help and support far beyond what would be ideally expected from a doctoral thesis director. A full chapter could be written about the many different ways he has guided me, not only during this doctoral thesis, but from the very beginning of our relationship (back in 2007).

Special thanks to Prof. Jaume Vives for co-directing this doctoral thesis and for all the hard work and effort he has put into it.

Special gratitude goes also to my dad, my mom, and Natza for their uninterested financial help when I've been out of cash. I'm in debt with them.

There are many more people who have helped me in this journey, but I could not mention them all. I really hope they know how grateful I am for their support.

Finally, I would like to thank Prof. K. Fee for his uninterrupted support and loyalty in the best and worst moments of this adventure.

7 References

Abel, U., & Koch, A. (1999). The role of randomization in clinical studies: myths and beliefs. *Journal of clinical epidemiology*, *52*(6), 487–497. doi:10.1016/S0895-4356(99)00041-4

American Psychological Association, National Council on Measurement in Education, & American Educational Research Association. (1999). *Standards for educational and psychological testing* (2nd ed.). American Educational Research Association.

Benson, K., & Hartz, A. J. (2000). A Comparison of observational studies and randomized, controlled trials. *New England Journal of Medicine*, *342*(25), 1878–1886. doi:10.1056/NEJM200006223422506

Black, N. (1996). Why we need observational studies to evaluate the effectiveness of health care. *British Medical Journal*, *312*(7040), 1215–1218. doi:10.1136/bmj.312.7040.1215

Bluhm, R. (2009). Some observations on “observational” research. *Perspectives in Biology and Medicine*, *52*(2), 252–63. doi:10.1353/pbm.0.0076

Byar, D. P., Simon, R. M., Friedewald, W. T., Schlesselman, J. J., DeMets, D. L., Ellenberg, J. H., … Ware, J. H. (1976). Randomized clinical trials. Perspectives on some recent ideas. *The New England Journal of Medicine*, *295*(2), 74. doi:10.1056/NEJM197607082950204

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.

Concato, J., Shah, N., & Horwitz, R. I. (2000). Randomized, controlled trials, observational studies, and the hierarchy of research designs. *New England Journal of Medicine*, *342*(25), 1887–1892. doi:10.1056/NEJM200006223422507

Cook, D. J., Sackett, D. L., & Spitzer, W. O. (1995). Methodologic guidelines for systematic reviews of randomized control trials in health care from the Potsdam Consultation on Meta-Analysis. *Journal of Clinical Epidemiology*, *48*(1), 167. doi:10.1016/0895-4356(94)00172-M

Costa, B. R. da, Cevallos, M., Altman, D. G., Rutjes, A. W. S., & Egger, M. (2011). Uses and misuses of the STROBE statement: bibliographic study. *BMJ Open*, *1*(1). doi:10.1136/bmjopen-2010-000048

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334.

Crowther, M. A., & Cook, D. J. (2007). Trials and tribulations of systematic reviews and meta-analyses. *ASH Education Program Book*, *2007*(1), 493–497. doi:10.1182/asheducation-2007.1.493

Deeks, J. J., Dinnes, J., D'Amico, R., Sowden, A. J., Sakarovitch, C., Song, F., … Altman, D. G. (2003). Evaluating non-randomised intervention studies. *Health Technology Assessment*, *7*(27), 1–173. doi:10.3310/hta7270

Detsky, A. S., Naylor, C. D., O'Rourke, K., McGeer, A. J., & L'Abbé, K. A. (1992). Incorporating variations in the quality of individual randomized trials into meta-analysis. *Journal of Clinical Epidemiology*, *45*(3), 255–265. doi:16/0895-4356(92)90085-2

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. doi:10.1136/jech.52.6.377

Dreier, M., Borutta, B., Stahmeyer, J., Krauth, C., & Walter, U. (2010). Comparison of tools for assessing the methodological quality of primary and secondary studies in health technology assessment reports in Germany. *GMS Health Technology Assessment*, *6*. doi:10.3205/hta000085

Feinstein, A. R. (1984). Current problems and future challenges in randomized clinical trials. *Circulation*, *70*(5), 767–774. doi:10.1161/01.CIR.70.5.767

Funai, E. F., Rosenbush, E. J., Lee, M.-J., & Del Priore, G. (2001). Distribution of study designs in four major US journals of obstetrics and gynecology. *Gynecologic and Obstetric Investigation*, *51*(1), 8–11. doi:10.1159/000052882

Greene, T. (2000). Are observational studies "just as effective" as randomized clinical trials? *Blood Purification*, *18*(4), 317–322. doi:10.1159/000014455

Greenland, S. (1994). Quality scores are useless and potentially misleading: reply to "Re: A critical look at some popular analytic methods". *American Journal of Epidemiology*, *140*(3), 300–301.

Groenwold, R. H. H., Nelson, D. B., Nichol, K. L., Hoes, A. W., & Hak, E. (2010). Sensitivity analyses to estimate the potential impact of unmeasured confounding in causal research. *International Journal of Epidemiology*, *39*(1), 107 –117. doi:10.1093/ije/dyp332

Harbour, R., & Miller, J. (2001). A new system for grading recommendations in evidence based guidelines. *British Medical Journal, 323*(7308), 334–336. doi:10.1136/bmj.323.7308.334

Hartling, L., Bond, K., Harvey, K., Santaguida, P. L., Viswanathan, M., & Dryden, D. M. (2010). *Developing and testing a tool for the classification of study designs in systematic reviews of interventions and exposures* (Methods Research Report No. 11-EHC-007). Agency for Healthcare Research and Quality. Retrieved from http://effectivehealthcare.ahrq.gov/

Hartz, A., Benson, K., Glaser, J., Bentler, S., & Bhandari, M. (2003). Assessing observational studies of spinal fusion and chemonucleolysis. *Spine*, *28*(19), 2268–2275.

Herbison, P., Haysmith, J., & Gillespie, W. (2006). Adjustment of meta-analyses on the basis of quality scores should be abandoned. *Journal of Clinical Epidemiology*, *59*(12), 1249.e1–1249.e11. doi:10.1016/j.jclinepi.2006.03.008

Higgins, J. P. T., & Green, S. (2011). *Cochrane handbook for systematic reviews of interventions*. New York: John Wiley & Sons Inc.

Jadad, A. R. (2000). Systematic reviews and meta-analyses on treatment of asthma: critical evaluation. *British Medical Journal*, *320*(7234), 537–540. doi:10.1136/bmj.320.7234.537

Jadad, A. R., Moher, D., & Klassen, T. P. (1998). Guides for reading and interpreting systematic reviews: II. How did the authors find the studies and assess their quality? *Archives of Pediatrics & Adolescent Medicine, 152*(8), 812. doi:10.1001/archpedi.152.8.812

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. doi:10.1136/bmj.323.7303.42

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. doi:10.1001/jama.282.11.1054

Kane, R. L. (1997). Approaching the outcomes question. In Kane, R. L. (Ed.), *Understanding health care outcomes research* (pp. 1–15). Jones & Bartlett Learning.

Khan, K. S., Daya, S., & Jadad, A. R. (1996). The importance of quality of primary studies in producing unbiased systematic reviews. *Archives of Internal Medicine, 156*(6), 661–666. doi:10.1001/archinte.1996.00440060089011

MacLehose, R., Reeves, B., Harvey, I., Sheldon, T., Russell, I., & Black, A. (2000). A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies. *Health Technology Assessment*, *4*(34), 1–154. doi:10.3310/hta4340

Manchikanti, L. (2008). Evidence-based medicine, systematic reviews, and guidelines in interventional pain management, part I: introduction and general considerations. *Pain physician*, *11*(2), 161–186.

Manchikanti, L., Benyamin, R. M., Helm, S., & Hirsch, J. A. (2009a). Evidence-based medicine, systematic reviews, and guidelines in interventional pain management: part 3: systematic reviews and meta-analyses of randomized trials. *Pain Physician*, *12*(1), 35.

Manchikanti, L., Datta, S., Smith, H. S., & Hirsch, J. A. (2009b). Evidence-based medicine, systematic reviews, and guidelines in interventional pain management: part 6. Systematic reviews and meta-analyses of observational studies. *Pain Physician*, *12*(5), 819–850.

Manchikanti, L., Hirsch, J. A., & Smith, H. S. (2008). Evidence-based medicine, systematic reviews, and guidelines in interventional pain management: Part 2: Randomized controlled trials. *Pain physician*, *11*(6), 717–773.

Manchikanti, L., Singh, V., Smith, H. S., & Hirsch, J. A. (2009c). Evidence-based medicine, systematic reviews, and guidelines in interventional pain management: part 4: observational studies. *Pain Physician*, *12*(1), 73–108.

Mann, C. J. (2003). Observational research methods. Research design II: cohort, cross sectional, and case-control studies. *Emergency Medicine Journal*, *20*(1), 54–60. doi:10.1136/emj.20.1.54

Marín-Martínez, F., Sánchez-Meca, J., Huedo, T., & Fernández, I. (2007). Meta-análisis: ¿Dónde estamos y hacia dónde vamos? In *Psicología y Ciencias Afines en los Albores del Siglo XXI (Homenaje al Profesor Sánchez Bruno).* (pp. 87–102). Granada: Grupo Editorial Universitario.

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. doi:10.1016/j.eurpsy.2005.03.011

McAlister, F. A., Clark, H. D., Van Walraven, C., Straus, S. E., Lawson, F. M., Moher, D., & Mulrow, C. D. (1999). The medical review article revisited: has the science improved? *Annals of Internal Medicine*, *131*(12), 947.

McKimmie, T., & Szurmak, J. (2002). Beyond grey literature: How grey questions can drive research. *Journal of Agricultural & Food Information*, *4*(2), 71–79. doi:10.1300/J108v04n02_06

McMahon, A. D. (2003). Approaches to combat with confounding by indication in observational studies of intended drug effects. *Pharmacoepidemiology and Drug Safety*, *12*(7), 551–558. doi:10.1002/pds.883

Moher, D., Cook, D. J., Eastwood, S., Olkin, I., Rennie, D., & Stroup, D. F. (1999). Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. *The Lancet*, *354*(9193), 1896–1900. doi:10.1016/S0140-6736(99)04149-5

Moher, D., Jadad, A. R., & Tugwell, P. (1996). Assessing the quality of randomized controlled trials: current issues and future directions. *International Journal of Technology Assessment in Health Care*, *12*(02), 195–208. doi:10.1017/S0266462300009570

Montero, I., & León, O. G. (2007). A guide for naming research studies in psychology. *International Journal of Clinical and Health Psychology*, *7*(3), 847–862.

Normand, S.-L. T. (2005). Readers guide to critical appraisal of cohort studies: 3. Analytical strategies to reduce confounding. *British Medical Journal*, *330*, 1021–1023. doi:10.1136/bmj.330.7498.1021

Petrie, A., Bulman, J. S., & Osborn, J. F. (2003). Further statistics in dentistry Part 8: Systematic reviews and meta-analyses. *British Dental Journal*, *194*(2), 73–78. doi:10.1038/sj.bdj.4809877

Petticrew, M., Song, F., Wilson, P., & Wright, K. (1999). Quality-assessed reviews of health care interventions and the database of abstracts of reviews of effectiveness (DARE). *International Journal of Technology Assessment in Health Care*, *15*(4), 671–678.

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 published medical education research. *Journal of the American Medical Association*, *298*(9), 1002–1009. doi:10.1001/jama.298.9.1002

Rosenbaum, P. R. (1991). Discussing hidden bias in observational studies. *Annals of Internal Medicine*, *115*(11), 901 –905. doi:10.1059/0003-4819-115-11-901

Sánchez-Meca, J. (2010). Cómo realizar una revisión sistemática y un meta-análisis. *Aula abierta*, *38*(2), 53–64.

Sánchez-Meca, J., & Botella, J. (2010). Revisiones sistemáticas y meta-análisis: Herramientas para la práctica profesional. *Papeles del psicólogo: revista del Colegio Oficial de Psicólogos*, *31*(1), 7–17.

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. doi:10.1093/ije/dym018

Scales Jr, C. D., Norris, R. D., Peterson, B. L., Preminger, G. M., & Dahm, P. (2005). Clinical research and statistical methods in the urology literature. *The Journal of Urology, 174*(4, Part 1), 1374–1379. doi:10.1097/01.ju.0000173640.91654.b5

Shamliyan, T. A., Kane, R. L., Ansari, M. T., Raman, G., Berkman, N. D., Grant, M., … Tsouros, S. (2010). *Development quality criteria to evaluate nontherapeutic studies of incidence, prevalence, or risk factors of chronic diseases: pilot study of new checklists* (Methods Research Report No. 11-EHC008-EF). Agency for Healthcare Research and Quality.

Shikata, S., Nakayama, T., Noguchi, Y., Taji, Y., & Yamagishi, H. (2006). Comparison of effects in randomized controlled trials with observational studies in digestive surgery. *Annals of Surgery, 244*(5), 668–676. doi:10.1097/01.sla.0000225356.04304.bc

Shrier, I., Boivin, J. F., Steele, R. J., Platt, R. W., Furlan, A., Kakuma, R., … Rossignol, M. (2007). Should meta-analyses of interventions include observational studies in addition to randomized controlled trials? A critical examination of underlying principles. *American Journal of Epidemiology, 166*(10), 1203–1209. doi:10.1093/aje/kwm189

Thompson, S., Ekelund, U., Jebb, S., Lindroos, A. K., Mander, A., Sharp, S., … Wilks, D. (2010). A proposed method of bias adjustment for meta-analyses of published observational studies. *International Journal of Epidemiology, 40*, 765–777. doi:10.1093/ije/dyq248

Vandenbroucke, J. P., Von Elm, E., Altman, D. G., Gøtzsche, P. C., Mulrow, C. D., Pocock, S. J., … Egger, M. (2007). Strengthening the reporting of observational studies in epidemiology (STROBE): explanation and elaboration. *Epidemiology, 18*(6), 805–835. doi:10.1097/EDE.0b013e3181577511

Viswanathan, M., & Berkman, N. D. (2012). Development of the RTI item bank on risk of bias and precision of observational studies. *Journal of Clinical Epidemiology, 65*(2), 163–178. doi:10.1016/j.jclinepi.2011.05.008

Von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C., & Vandenbroucke, J. P. (2007). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Preventive Medicine, 45*(4), 247–251. doi:10.1016/j.ypmed.2007.08.012

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. doi:10.1177/1049731508317278

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). Rockville, MD: Agency for Healthcare Research and Quality.

Whiting, P., Rutjes, A. W., Reitsma, J. B., Bossuyt, P. M., & Kleijnen, J. (2003). The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Medical Research Methodology*, *3*(1), 25. doi:10.1186/1471-2288-3-25

8
Annexes

Annex 1: Manual of the Data Extraction Form (Article 1)

MANUAL DEL FORMULARIO DATA EXTRACTOR

Identificación de la Herramienta

Código de Identificación RefWorks

Es el código con que está identificado el documento en la base de datos RefWorks. En el caso que un mismo documento tenga más de una herramienta se ingresarán en registros diferentes y se utilizará como identificador el identificador +/-1.

Título del Documento

Las primeras palabras del título para poderlo localizar con facilidad.

Nombre de la Herramienta

Nombre/acrónimo de la herramienta si los autores la han nombrado o apellido del primer autor seguido del año de publicación. Si un mismo autor presenta diferentes herramientas para diferentes tipos de estudio se añadirá '_Coh', '_CC', '_CS' para estudios de cohortes, de casos y controles y transversales específicamente.

Primer Autor

Indicar en formato APA. En el caso de desconocerse (documento de Internet) marcar este campo con una arroba [@].

Año de Publicación

Indicar el año con cuatro cifras. En el caso de herramientas encontradas en Internet de las que no se tenga el año de 'publicación', indicar con '0000'.

Tipo de Estudio al que se aplica la herramienta

Estudios observacionales.(no específic.). Marcaremos esta opción como 'Sí' sólo cuando los autores utilicen este término sin especificar más detalles.

Estudios de cohortes. Marcaremos esta opción como 'Sí' sólo cuando los autores indiquen que la herramienta es aplicable a este tipo de estudios o que hayan preguntas claramente dirigidas a estos diseños (por ejemplo preguntas sobre el seguimiento o la comparabilidad de las cohortes). A no ser que se deduzca lo contrario, también se marcará esta opción como 'Sí' si se habla de estudios longitudinales.

Estudios de casos i controles. Marcaremos esta opción como 'Sí' sólo cuando los autores indiquen que la herramienta es aplicable a este tipo de estudios o que hayan preguntas claramente dirigidas a estos diseños (por ejemplo preguntas sobre la similitud de los casos y los controles).

Estudios transversales analíticos. Marcaremos esta opción como 'Sí' sólo cuando los autores indiquen que la herramienta es aplicable a este tipo de estudios o que hayan preguntas claramente dirigidas a estos diseños.

Otros estudios no experimentales. Marcaremos esta opción como 'Sí' sólo cuando los autores indiquen que la herramienta es aplicable a cualquier otro tipo de diseño no experimental, o que haya preguntas claramente dirigidas a estos diseños. Si hablan de estudios observacionales pero lo matizan indicando los estudios de cohortes, casos y controles y transversales no se marcará esta opción como 'Sí'.

Estudios experimentales. Marcaremos esta opción como 'Sí' sólo cuando los autores indiquen que la herramienta es aplicable a cualquier otro tipo de diseño no experimental, o que haya preguntas claramente dirigidas a estos diseños.

Tipo de Herramienta

Checklist. Se considerará un checklist cuando todos los ítems sean de respuesta categórica (ordinal o nominal).

Escala. Se considerará una escala cuando todos los ítems sean de respuesta cuantitativa (escala métrica).

Mixto. Se considerará un sistema mixto aquella herramienta que tenga ítems propios de un checklist y también ítems propios de una escala.

Valoración Global

Indicar si los autores contemplan explícitamente la existencia de una valoración global y si en su cálculo existe algún tipo de ponderación.

Número de Ítems

Indicar el número de ítems. Se contabilizarán como ítems sólo los enunciados con opción de respuesta cerrada. Si un ítem consta de varios enunciados pero sólo permite dar una respuesta sólo se contará como un ítem. Si hay ítems de respuesta abierta no se contabilizarán.

Desarrollo de la Herramienta

Discusión del Concepto de Calidad

Indicar como ‘Sí’ si existe una discusión acerca del concepto de calidad y si se indican tipos o dominios de calidad (por ejemplo validez interna) que se pretenden evaluar con la herramienta. Si sólo se da por encima una definición de calidad o si sólo se diferencia por ejemplo validez interna de la externa, entonces se valorará como ‘Parcial’. También se valorará como ‘Parcial’ si en la documentación no se discute el concepto de calidad pero en la herramienta se desglosa por ejemplo validez interna de validez externa.

Adaptación/Modificación de otra herramienta

Marcar como ‘Sí’ cuando los autores citan que se han basado en una o varias herramientas ya existentes o si presentan su herramienta como una adaptación.

Procedimiento empírico de selección de ítems

Marcar esta opción cuando los ítems se hayan desarrollado a partir de estudios empíricos con expertos utilizando por ejemplo técnicas delphi, focus group o un proceso iterativo de selección de ítems.

Prueba Piloto

Indicar si se realiza una prueba piloto de la herramienta.

Fiabilidad entre Jueces

Marcar como ‘Sí’ si se evalúa la fiabilidad entre jueces y se indican los resultados. Si se discute sobre la fiabilidad de la herramienta pero no se presentan resultados marcar como ‘Parcial’.

Validez

Marcar como ‘Sí’ si se evalúan varias dimensiones de validez (contenido, constructo, concurrente…) y como ‘Parcial’ si sólo se evalúa un de las dimensiones de validez.

Contenido de la Herramienta

Definición Explícita de los Criterios de Selección

Marcar como ‘Sí’ si la herramienta cubre, ya sea mediante una pregunta concreta o mediante varias preguntas, alguno los puntos de cada dominio.

Definición explícita de los criterios de selección de los participantes. (Coh, Trans) y de tanto los casos como los controles (CC).

Los criterios de selección se han aplicado de la misma manera a todos los grupos.

Representatividad de la Muestra

Justificación de la representatividad de la muestra o selección aleatoria. No hay diferencia entre los participantes y los no participantes.

Comparabilidad de los Grupos

Marcar como ‘Sí’ si la herramienta cubre, ya sea mediante una pregunta concreta o mediante varias preguntas, todos los puntos del. En el caso de cubrir sólo alguno de los puntos, marcar este dominio como ‘Parcial’.

Los grupos de participantes proceden de la misma población o de poblaciones comparables.

(CC) Los controles son comparables a los casos excepto en la condición de interés.

Métodos de Medida

La herramienta obliga a valorar los instrumentos de medida utilizados en el estudio.

Fuentes de Sesgo controladas por Diseño

Se controlan posibles fuentes de sesgo durante la recogida de datos tales como sesgo de recuerdo, sesgo del entrevistador o pérdida selectiva de participantes.

Métodos de Control de la Confusión mediante el Diseño

Se aplican técnicas de control para aumentar la comparabilidad de los grupos de estudio respecto a potenciales factores de confusión.

Métodos de Control de la Confusión mediante el Análisis Estadístico

Se han controlado las variables potencialmente confundidoras en el análisis estadístico.

Análisis de las Pérdidas de Seguimiento

(Coh) Los grupos que se comparan (expuestos y no expuestos) no difieren en cuanto a los sujetos que abandonan o de los que se pierde el seguimiento.

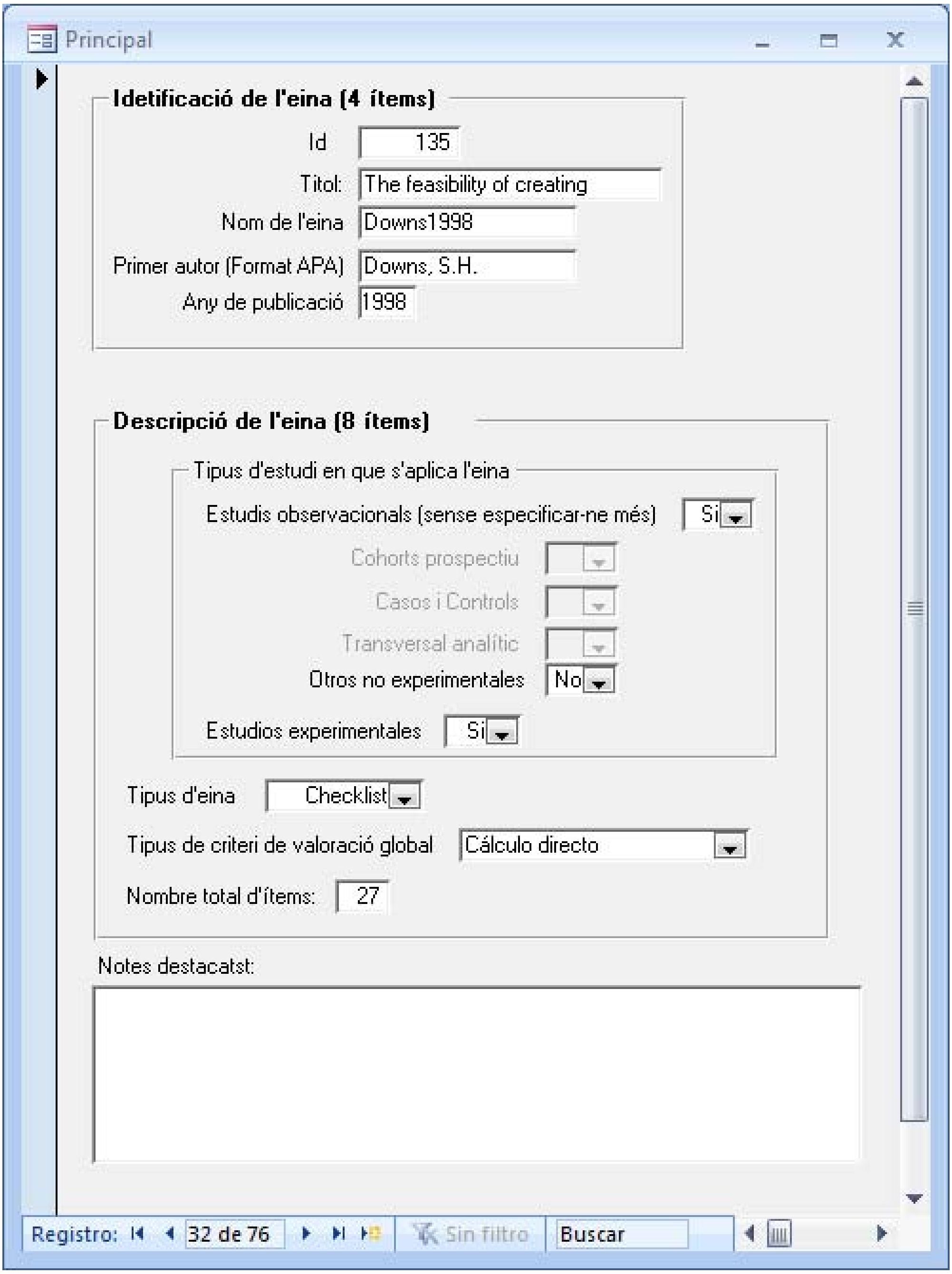
Análisis de las Ausencias de Datos

La ausencia de datos afecta de forma igual a los sujetos de los diferentes grupos.

Conflicto de intereses

Considera el origen de la financiación y posibles conflictos de intereses.

Annex 2: Data Extraction Form Screenshots (Article 1)



Principal

Idetificació de l'eina (4 ítems)

Id: 135

Titol: The feasibility of creating

Nom de l'eina: Downs1998

Primer autor (Format APA): Downs, S.H.

Any de publicació: 1998

Descripció de l'eina (8 ítems)

Tipus d'estudi en que s'aplica l'eina

Estudis observacionals (sense especificar-ne més): Si

Cohorts prospectiu:

Casos i Controls:

Transversal analític:

Otros no experimentales: No

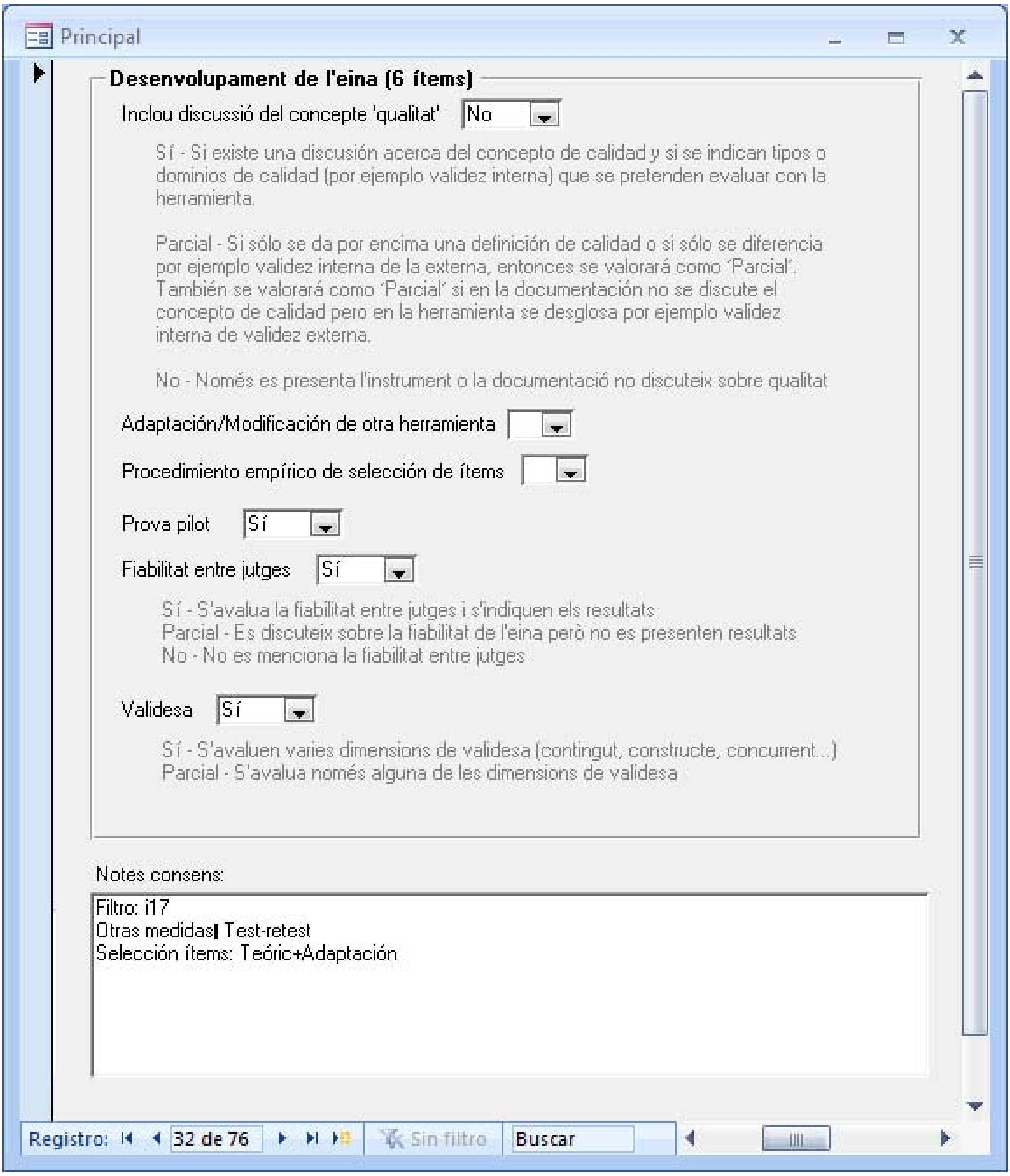
Estudios experimentales: Si

Tipus d'eina: Checklist

Tipus de criteri de valoració global: Cálculo directo

Nombre total d'ítems: 27

Notes destacatst:



Desenvolupament de l'eina (6 ítems)

Inclou discussió del concepte 'qualitat' No

Sí - Si existe una discusión acerca del concepto de calidad y si se indican tipos o dominios de calidad (por ejemplo validez interna) que se pretenden evaluar con la herramienta.

Parcial - Si sólo se da por encima una definición de calidad o si sólo se diferencia por ejemplo validez interna de la externa, entonces se valorará como 'Parcial'. También se valorará como 'Parcial' si en la documentación no se discute el concepto de calidad pero en la herramienta se desglosa por ejemplo validez interna de validez externa.

No - Només es presenta l'instrument o la documentació no discuteix sobre qualitat

Adaptación/Modificación de otra herramienta

Procedimiento empírico de selección de ítems

Prova pilot Sí

Fiabilitat entre jutges Sí

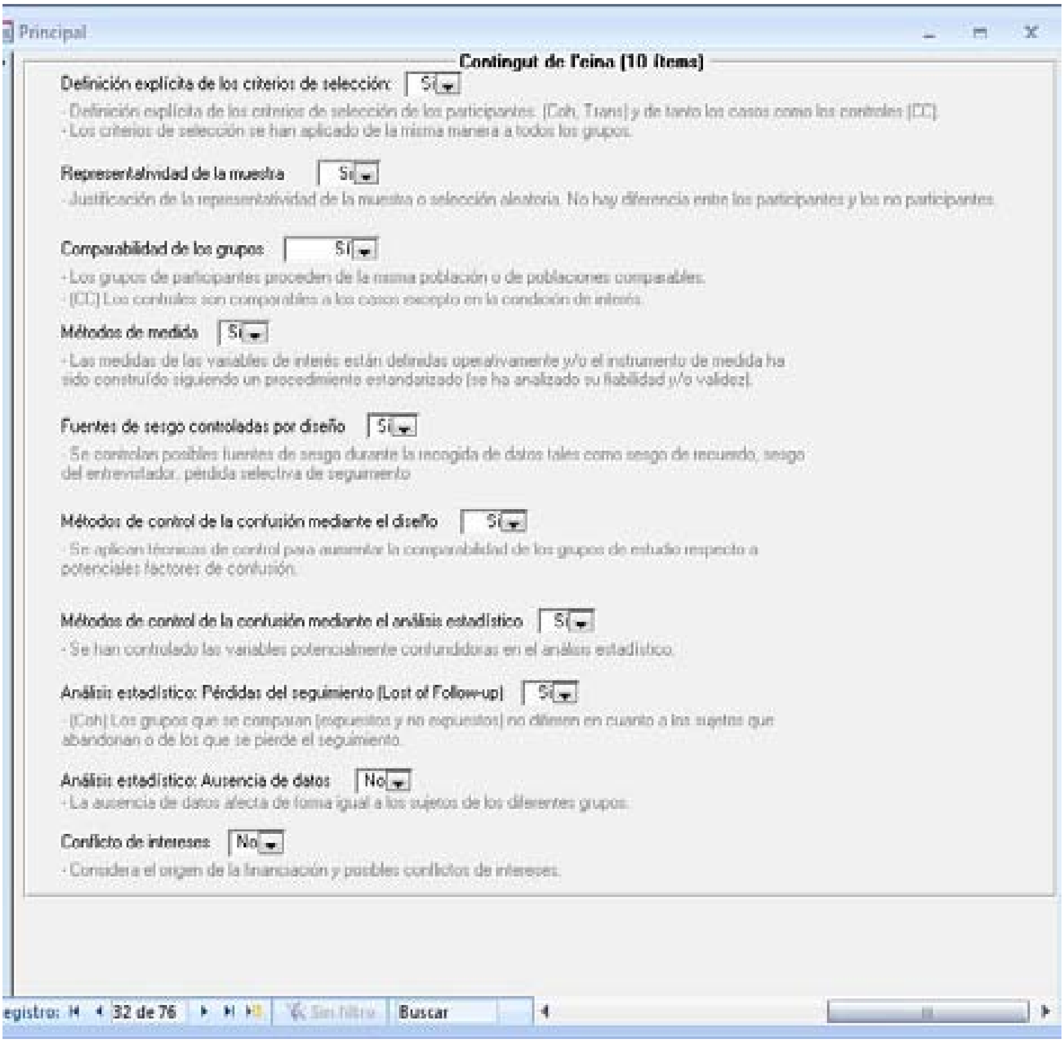
Sí - S'avalua la fiabilitat entre jutges i s'indiquen els resultats
Parcial - Es discuteix sobre la fiabilitat de l'eina però no es presenten resultats
No - No es menciona la fiabilitat entre jutges

Validesa Sí

Sí - S'avaluen varies dimensions de validesa (contingut, constructe, concurrent...)
Parcial - S'avalua només alguna de les dimensions de validesa

Notes consens:

Filtro: i17
Otras medidas: Test-retest
Selección ítems: Teóric+Adaptación



Contingut de l'eina (10 ítems)

Definición explícita de los criterios de selección: Sí

- Definición explícita de los criterios de selección de los participantes. (Coh, Trans) y de tanto los casos como los controles (CC).
- Los criterios de selección se han aplicado de la misma manera a todos los grupos.

Representatividad de la muestra Sí

- Justificación de la representatividad de la muestra o selección aleatoria. No hay diferencia entre los participantes y los no participantes.

Comparabilidad de los grupos Sí

- Los grupos de participantes proceden de la misma población o de poblaciones comparables.
- (CC) Los controles son comparables a los casos excepto en la condición de interés.

Métodos de medida Sí

- Las medidas de las variables de interés están definidas operativamente y/o el instrumento de medida ha sido construído siguiendo un procedimiento estandarizado (se ha analizado su fiabilidad y/o validez).

Fuentes de sesgo controladas por diseño Sí

- Se controlan posibles fuentes de sesgo durante la recogida de datos tales como sesgo de recuerdo, sesgo del entrevistador, pérdida selectiva de seguimento

Métodos de control de la confusión mediante el diseño Sí

- Se aplican técnicas de control para aumentar la comparabilidad de los grupos de estudio respecto a potenciales factores de confusión.

Métodos de control de la confusión mediante el análisis estadístico Sí

- Se han controlado las variables potencialmente confundidoras en el análisis estadístico.

Análisis estadístico: Pérdidas del seguimiento (Lost of Follow-up) Sí

- (Coh) Los grupos que se comparan (expuestos y no expuestos) no difieren en cuanto a los sujetos que abandonan o de los que se pierde el seguimiento.

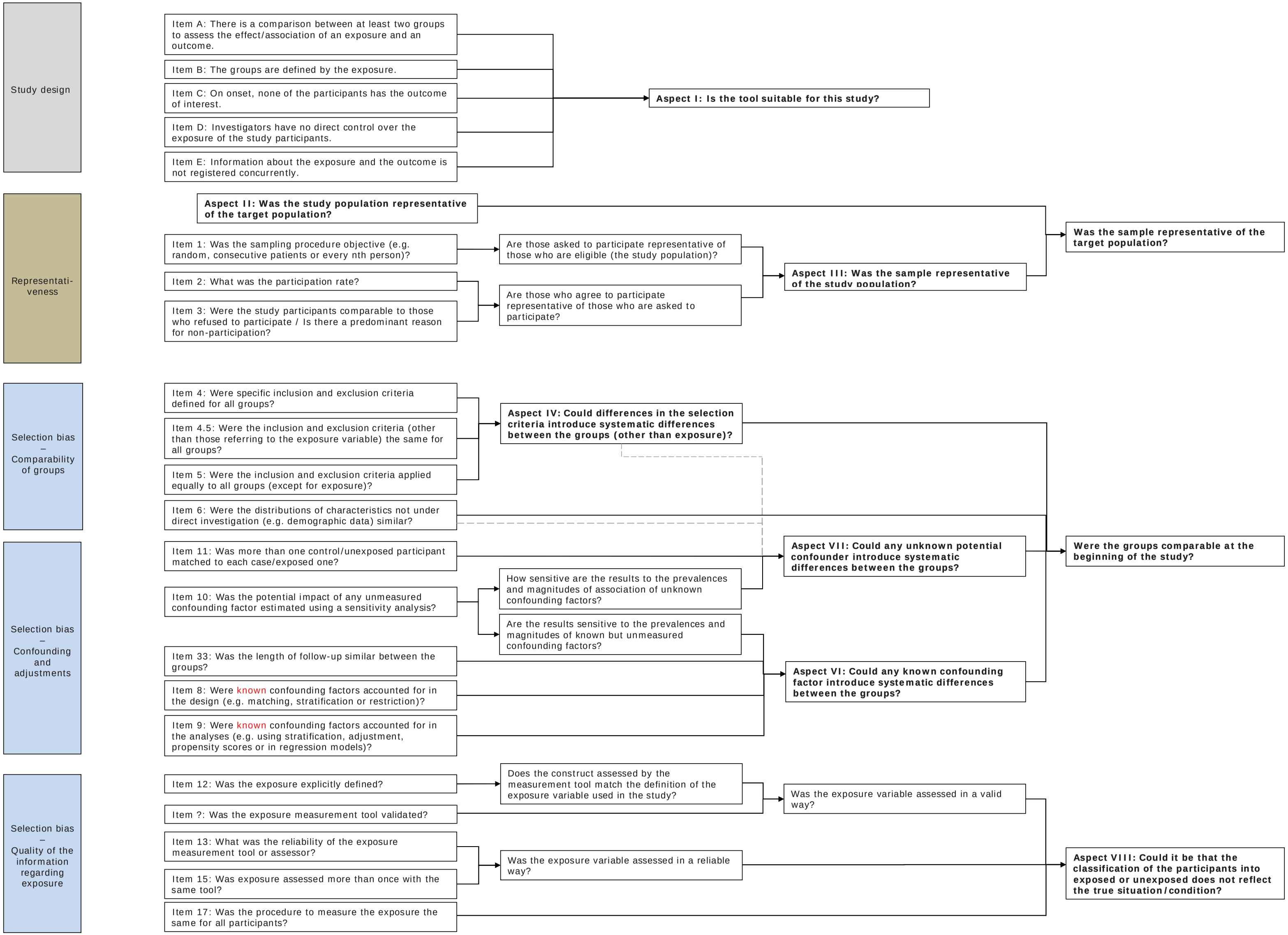
Análisis estadístico: Ausencia de datos No

- La ausencia de datos afecta de forma igual a los sujetos de los diferentes grupos.

Conflicto de intereses No

- Considera el origen de la financiación y posibles conflictos de intereses.

Annex 3: Diagram of the tool’s draft



Study design

Item A: There is a comparison between at least two groups to assess the effect/association of an exposure and an outcome.

Item B: The groups are defined by the exposure.

Item C: On onset, none of the participants has the outcome of interest.

Item D: Investigators have no direct control over the exposure of the study participants.

Item E: Information about the exposure and the outcome is not registered concurrently.

Aspect I: Is the tool suitable for this study?

Representati-veness

Aspect II: Was the study population representative of the target population?

Item 1: Was the sampling procedure objective (e.g. random, consecutive patients or every nth person)?

Item 2: What was the participation rate?

Item 3: Were the study participants comparable to those who refused to participate / Is there a predominant reason for non-participation?

Are those asked to participate representative of those who are eligible (the study population)?

Are those who agree to participate representative of those who are asked to participate?

Aspect III: Was the sample representative of the study population?

Was the sample representative of the target population?

Selection bias – Comparability of groups

Item 4: Were specific inclusion and exclusion criteria defined for all groups?

Item 4.5: Were the inclusion and exclusion criteria (other than those referring to the exposure variable) the same for all groups?

Item 5: Were the inclusion and exclusion criteria applied equally to all groups (except for exposure)?

Aspect IV: Could differences in the selection criteria introduce systematic differences between the groups (other than exposure)?

Item 6: Were the distributions of characteristics not under direct investigation (e.g. demographic data) similar?

Selection bias – Confounding and adjustments

Item 11: Was more than one control/unexposed participant matched to each case/exposed one?

Item 10: Was the potential impact of any unmeasured confounding factor estimated using a sensitivity analysis?

How sensitive are the results to the prevalences and magnitudes of association of unknown confounding factors?

Are the results sensitive to the prevalences and magnitudes of known but unmeasured confounding factors?

Aspect VII: Could any unknown potential confounder introduce systematic differences between the groups?

Item 33: Was the length of follow-up similar between the groups?

Item 8: Were known confounding factors accounted for in the design (e.g. matching, stratification or restriction)?

Item 9: Were known confounding factors accounted for in the analyses (e.g. using stratification, adjustment, propensity scores or in regression models)?

Aspect VI: Could any known confounding factor introduce systematic differences between the groups?

Were the groups comparable at the beginning of the study?

Selection bias – Quality of the information regarding exposure

Item 12: Was the exposure explicitly defined?

Item ?: Was the exposure measurement tool validated?

Does the construct assessed by the measurement tool match the definition of the exposure variable used in the study?

Was the exposure variable assessed in a valid way?

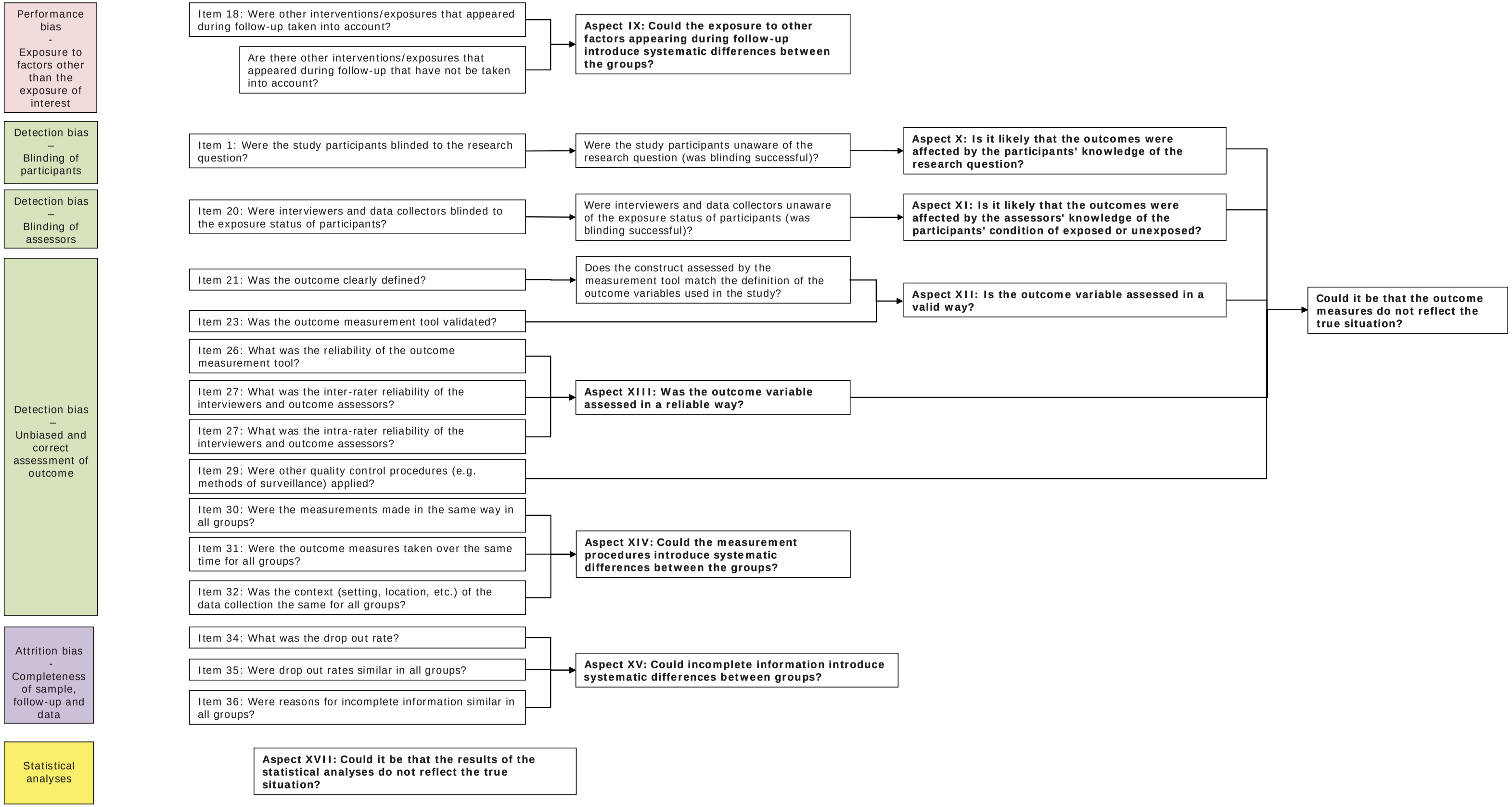
Item 13: What was the reliability of the exposure measurement tool or assessor?

Item 15: Was exposure assessed more than once with the same tool?

Was the exposure variable assessed in a reliable way?

Item 17: Was the procedure to measure the exposure the same for all participants?

Aspect VIII: Could it be that the classification of the participants into exposed or unexposed does not reflect the true situation/condition?



Performance bias - Exposure to factors other than the exposure of interest

Item 18: Were other interventions/exposures that appeared during follow-up taken into account?

Are there other interventions/exposures that appeared during follow-up that have not be taken into account?

Aspect IX: Could the exposure to other factors appearing during follow-up introduce systematic differences between the groups?

Detection bias – Blinding of participants

Item 1: Were the study participants blinded to the research question?

Were the study participants unaware of the research question (was blinding successful)?

Aspect X: Is it likely that the outcomes were affected by the participants' knowledge of the research question?

Detection bias – Blinding of assessors

Item 20: Were interviewers and data collectors blinded to the exposure status of participants?

Were interviewers and data collectors unaware of the exposure status of participants (was blinding successful)?

Aspect XI: Is it likely that the outcomes were affected by the assessors' knowledge of the participants' condition of exposed or unexposed?

Detection bias – Unbiased and correct assessment of outcome

Item 21: Was the outcome clearly defined?

Does the construct assessed by the measurement tool match the definition of the outcome variables used in the study?

Item 23: Was the outcome measurement tool validated?

Aspect XII: Is the outcome variable assessed in a valid way?

Item 26: What was the reliability of the outcome measurement tool?

Item 27: What was the inter-rater reliability of the interviewers and outcome assessors?

Item 27: What was the intra-rater reliability of the interviewers and outcome assessors?

Aspect XIII: Was the outcome variable assessed in a reliable way?

Item 29: Were other quality control procedures (e.g. methods of surveillance) applied?

Could it be that the outcome measures do not reflect the true situation?

Item 30: Were the measurements made in the same way in all groups?

Item 31: Were the outcome measures taken over the same time for all groups?

Item 32: Was the context (setting, location, etc.) of the data collection the same for all groups?

Aspect XIV: Could the measurement procedures introduce systematic differences between the groups?

Attrition bias - Completeness of sample, follow-up and data

Item 34: What was the drop out rate?

Item 35: Were drop out rates similar in all groups?

Item 36: Were reasons for incomplete information similar in all groups?

Aspect XV: Could incomplete information introduce systematic differences between groups?

Statistical analyses

Aspect XVII: Could it be that the results of the statistical analyses do not reflect the true situation?

Annex 4: Manual of the Q-Coh

MANUAL OF THE QUALITY ASSESSMENT TOOL FOR COHORT STUDIES Q-COH

Alexander Jarde (A.Jarde@gmail.com)
Josep Maria Losilla
Jaume Vives

Introduction

In this manual for the quality assessment tool we will try to give as extensive information as possible in order to empower users in their usage of theQ-Coh, especially regarding the subjective evaluations of the aspects and the quality.

Brief explanation of the structure

The tool is structured around six domains, which stem from a common classification scheme for bias (Selection bias, Performance bias, Detection bias, Attrition bias, and Reporting bias), plus the representativeness of the sample. Additionally, and given the lack of agreement in the definition of cohort studies, the applicability of the tool to the study is also assessed at the beginning.

There are two types of items: those assessing objective data, and those requiring a judgment (inferences). The first ones are the objective component of the tool and are written so that there is a minimum processing of the information available in the study's report for answering them. Inferences introduce a subjective component into the tool, since the user has to make a judgment. To do so the user has to use the information gathered by the objective items and combine them considering the particularities of the topic under study. This manual is specially oriented to aid this judgment process.

Purpose and framework

The purpose of this tool is to identify those cohort studies with low quality and therefore potential source of bias in the meta-analysis. It is not meant to be exhaustive, since there are aspects of the study's quality which might be too complex and variable (depending on the topic under study) to be assessed precisely with a closed tool, as for example the assessment of statistical analyses. Therefore, this tool focuses on the more essential aspects to set an acceptable level of methodological quality that a study should have, priming specificity over sensitivity.

Scope of the construct to be measured

The construct to be measured by this tool, labeled as *methodological quality* (or just *quality*), refers to the degree to which the study employs procedures to guarantee that the comparability of the groups is maintained along the whole study (and/or controlled for in the analyses), that the measures and results are valid and reliable, and that the results can be extrapolated to the target population.

As we can see, the definition of quality bases on three points. On one hand, the sustained comparability of the groups (or the proper controls in the analysis) is needed to be able to assure that the differences found between the groups are not due to any other systematic difference between them. On the other hand, it is important to assure that the measures and results reflect the true situation, because it is useless to have comparable groups if what we measure is not what we intend to measure. Similarly, if the raw data is correct but inappropriate statistical analyses are applied to obtain the results, these will not reflect the true situation again. Finally, the groups may have remained comparable and the measures and results may reflect the true situation, but if the sample is not representative of the target population the results won’t be applicable to it. This is important in the context of systematic reviews and meta-analyses, because at this level the primary studies’ target populations have to be comparable to the review’s and/or meta-analysis’ target population. If the raw data (primary studies’ sample data) is not representative of the studies’ target population, neither will it be of the review’s and/or meta-analysis’ target population.

On the contrary, our construct does not include aspects related to the correctness or completeness of the studies’ reporting, nor is related to other aspects considered of good research practice, but that are not susceptible to introduce systematic differences between the groups compared in the studies (e.g. ethical committee’s approval, sample size/power calculation).

Items and inferences

Study design

Although this tool is said to be applicable only to cohort studies, this and related terms are not used consistently by authors and databases along the health sciences. So, the label given to a study's design is not a reliable information for knowing if this tool is applicable to it. Therefore, the first five items assess the characteristics that define what is considered as a cohort study by this tool:

Item A: Is there a comparison between at least two groups to assess the effect/ association of an exposure and an outcome?

This means that there have to be at least two measures of the effect/ association of an exposure and an outcome that can be compared. If there is only one measure of association, we could only say how prone those exposed are to have the outcome. If two measures of association are compared, we can say how much more prone those exposed are to have the outcome than those not exposed (relative risk). So, to be able to compute an effect size, a relationship between at least two groups is needed.

This differentiates what we understand as a cohort study from other prospective or longitudinal studies, where only one group is studied (e.g. prevalence studies) without making a comparison with another reference group.

To answer this item positively ('Yes'), the effect/association of the exposure and the outcome had to be estimated for at least two groups.

Item B: Are the groups defined by the exposure variable?

This means that the different groups are formed by aggregating subjects exposed and unexposed (or with different degrees of exposure). Subjects may be selected from different populations (each defined by being exposed or not) or from the same population (defined by other selection criteria) and then classified into exposed or unexposed.

This differentiates what we understand as a cohort study from case-control studies, where the participants' level in the outcome variable defines the groups.

To answer this item positively ('Yes'), the groups that are compared have to be defined by their different value in the exposure variable.

Item C: Has or could any of the participants have the outcome of interest on onset?

This means that at the moment when the exposure variable was assessed (for the first time), none of the participants should have the outcome of interest.

This differentiates what we understand as a cohort study from a series of cross-sectional studies. It does not necessarily exclude retrospective cohort studies, though, where the outcome may not be absent in all participants at the beginning of the *study*.

To answer this item negatively ('No'), none of the participants may have the outcome of interest at the time the exposure is measured for the first time or those who have the outcome of interest are excluded. If the outcome is defined by a change in the value of a variable (for example an increase or decrease in perceived anxiety) then this item should also be answered 'No'. Answer this item positively ('Yes') if at least some of the participants have or could have the outcome of interest at the time the exposure is measured for the first time.

Item D: Do investigators handle who is exposed or not?

This means that the investigators do not allocate the participants into exposed or not.

This differentiates what we understand as a cohort study from randomized controlled trials or quasi-experimental controlled trials (among others).

To answer this item negatively ('No'), the investigators should have no control over who is exposed or not (no assignment to exposure).

Item E: Is information about the exposure and the outcome of interest registered concurrently?

This means that there has to be a time lapse between the exposure measurement and the outcome measurement. If the outcome is defined by a change in the value of a variable (for example and increase or decrease in perceived anxiety) there might be a concurrent assessment of the exposure and the base line of the outcome variable.

This differentiates what we understand as a cohort study from cross-sectional studies that might ask for the exposure retrospectively at the moment of assessing the outcome.

To answer this item negatively (‘No’), the measures of exposure and outcome that are put into relation have to be registered at different time points (the exposure preceding the outcome, of course).

Inference 0. Is the tool suitable for this study?

When judging this aspect the user has to consider the answers given to the items A to E. These items screen for the characteristics that define a cohort study as understood here, so if the items A and B are not answered positively ('Yes') or if items C, D or E have not been answered negatively ('No'), then the study has a design not contemplated by this tool and therefore should be assessed with another quality assessment tool. This is so regardless if the study is labeled as a ‘cohort’ study by either the authors or the database classification.

Representativeness

Item 1. Have the study participants been selected using a randomized sampling procedure?
Item 2. Is the similarity between the selected group of subjects and the target population justified by the authors?
Item 3. Is there a predominant reason for refusing to participate at the beginning of the study?
Inference 1. Could the results be generalized from the sample to the target population?

Item 1. Have the study participants been selected using a randomized sampling procedure?

It is important that the randomized sampling procedure guarantees that all the subjects of the target population have a chance of being selected. Attention should be paid for not getting confused by quasi-randomized sampling procedures.

To answer this item positively ('Yes'), the study participants should have been selected at random from a complete list of the target population or, if they are randomly selected from natural groups of subjects (e.g. hospitals, schools, cities, etc.), these also should have been selected at random among all groups that form the target population. Answer this item negatively ('No') if the study participants or the groups from where they are selected have been chosen using a non-randomized procedure (e.g. by convenience), hindering that all subjects of the target population had a chance of being selected.

The information assessed by this item should be available if the study was reported following the STROBE statement, as it is required in the item 6a: Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection.

Why is this important? A randomized sampling procedure enhances the representativeness of the sample, since it reduces the possibility that a subgroup of the target population is selected due to a common characteristic (e.g. younger, highly motivated, or healthier participants).

Item 2. Is the similarity between the selected group of subjects and the target population justified by the authors?

If the sampling procedure was not completely randomized, it is important that the authors justify that the selected group of subjects is similar to the target population despite it has not been selected using a randomized procedure. So, if it is shown that the sample has a similar distribution than the target population on certain variables, as for example demographic, clinical or social characteristics, confidence arises that the sampling procedure has maintained the similarity between the sample and the target population on other variables as well.

To answer this item as ‘Yes, empirically’, the similarity of the selected sample with the target population has to be empirically stated (e.g. comparing them on demographic, clinical or social characteristics). If the authors discuss the sampling procedure and justify the representativeness of the selected group of subjects without any empirical comparison then this item should be answered as ‘Yes, verbally’. Answer this item negatively (‘No’) if the similarity between the selected groups of subjects and the target population is not justified or is stated as a limitation of the study.

The information assessed by this item should be available if the study was reported following the STROBE statement, as it is required in the item 21: Discuss the generalisability (external validity) of the study results.

Why is this important? For a sample to be representative of the target population it has to be similar in all important variables, because if the sample has something ‘special’ there will be less evidence that the results can be extrapolated to those without that ‘special’ trait.

Item 3. Is there a predominant reason for refusing to participate at the beginning of the study?

It is important to not confound those who refuse to enter the study with participants who entered the study and dropped out along the follow-up or who refuse to participate when the outcome is assessed.

To answer this item negatively (‘No/Irrelevant’), participants and those refusing to enter the study should be similar or the number of subjects who refuse to participate should be small enough to be considered irrelevant. Answer this item positively (‘Yes’) if non-participants have something in common or give similar reasons for refusing to participate in the study. If no information is reported about the quantity or the reasons and characteristics of those refusing to participate in the study, this item should be answered as ‘Not reported’.

The information assessed by this item should be available if the study was reported following the STROBE statement, as it is required in the item 21: Discuss the generalizability (external validity) of the study results.

Why is this important? The group of people invited to participate in the study may very well be representative of the target population (maybe due to a randomized selection procedure), but if those refusing to participate share a common trait, the external validity will be compromised.

Inference 1. Considering the responses given to the items 1 to 3, could the results be generalized from the sample to the target population?

For the sample to be representative of the target population, the selected group of subjects should have been selected using a randomized sampling procedure that guarantees that all the subjects of the target population had a chance of being selected, and there should be no predominant reason for refusing entering the study. If the sampling procedure was not completely randomized, then the similarity between the sample and the target population should be justified.

Comparability of the groups

Item 4. Were the inclusion and exclusion criteria explicitly defined for all groups?
Item 5. Were the same inclusion and/or exclusion criteria applied equally to all groups?
Item 6 (inference). Could differences in the selection criteria introduce systematic differences between the groups (other than exposure)?
Item 7. Were known confounding factors accounted for in the design or in the analysis?
Inference.2. Is bias between the groups avoided at the beginning of the study?

Item 4: Were the inclusion and exclusion criteria explicitly defined for all groups?

This refers to the criteria each subject has to fulfill in order to be suitable for the study.

To answer this item positively ('Yes') there should be a clear and operative definition of the inclusion and exclusion criteria for all groups. If the inclusion and exclusion criteria are not reported or are broad and unspecific then this item should be answered negatively ('No').

The information assessed by this item should be available if the study was reported following the STROBE statement, as it is required in the item 6a: Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up.

Why is this important? Linking the inclusion of subjects into the study and into one group or another to specifically defined criteria reduces the risk that bias is introduced during the selection process. Additionally, clearly defined criteria enhance the replicability of the study.

Item 5: Were the same inclusion and/or exclusion criteria applied equally to all groups?

To answer this item positively (‘Yes’), those sampled individuals who were willing to participate should have been applied the same inclusion and exclusion criteria (except those referring to the exposure variable) in the same way. Answer this item negatively (‘No’) if the selection criteria were not the same or were applied differently in each group (e.g. different screening tools). If no selection criteria are reported or if it was not reported how they were applied then this item should be answered as ‘Not reported’.

The information assessed by this item should be available if the study was reported following the STROBE statement, as it is required in the item 6a: Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up.

Why is this important? Inclusion and exclusion criteria define what subjects are included or not into the study and in what group (exposed or unexposed) they are allocated. Besides those criteria referring to the exposure variable, it is desirable that the other selection criteria are the same for all groups in order to enhance the comparability between them. If the selection criteria are not applied in the same way to the parent populations of each group, systematic differences may arise between the groups.

Item 6 (inference): Considering the responses given to the items 4 and 5, could differences in the selection criteria introduce systematic differences between the groups (other than exposure)?

If the selection criteria are not specifically defined, or if they are different for the different groups or are not equally applied between them, the user has to carefully consider in how far each of these threats to the comparability may introduce systematic differences between the groups.

In order to answer this inference as 'Unlikely' the selection criteria should be explicitly defined and be applied equally to all groups. If the selection criteria are not explicitly defined, if they are different for the different groups, or if they are not equally applied between them, then it has to be carefully considered in how far each of these threats to the comparability may introduce systematic differences between the groups.

Item 7: Were known confounding factors accounted for in the design or in the analysis?

A confounding factor is a variable that has some effect on the outcome, and that is correlated with the exposure but without being affected by it (e.g. not be and intermediate step in the causal pathway between the exposure and the outcome)[1]. A variable is a confounding factor depending on the study question. Usually there are multiple risk factors for the disease of interest but researchers typically focus on the causal effect of only one of them. This factor is then the exposure, while the others are the confounding factors[2]. Confounding occurs when selection bias gives rise to imbalances between exposed and unexposed groups on confounding factors[3].

There are several ways of taking a confounding factor into account in the design: matching, stratification or restriction. The objective of these three procedures is to make the groups comparable on known confounding factors, either assuring a similar distribution of them (matching and stratification) or maintaining them constant (restriction).

Although if a confounding variable has not been accounted for in the design, it can still be controlled during the data analysis if it has been measured, for example using stratification, adjustment, propensity scores or in regression models.

To answer this item positively (‘Yes’), all the important confounding variables should have been accounted for in the design or in the statistical analyses. If some but not all the important confounding variables have been accounted for in the design or in the statistical analyses then answer this item as ‘Partially’. Answer this item negatively (‘No’) if none of the

[1] McNamee, R. (2003). Confounding and confounders. Occupational and Environmental Medicine, 60, 227–234. doi:10.1136/oem.60.3.227

[2] McNamee, R. (2005). Regression modelling and other methods to control confounding. Occupational and Environmental Medicine, 62(7), 500 –506. doi:10.1136/oem.2002.001115

[3] Higgins, J. P. T., & Green, S. (2011). Cochrane handbook for systematic reviews of interventions. New York: John Wiley & Sons Inc.

important confounding variables have been accounted for in the design or in the statistical analysis.

The information assessed by this item should be available if the study was reported following the STROBE statement, as it is required in the items 6b (For matched studies, give matching criteria and number of exposed and unexposed), 12a (Describe all statistical methods, including those used to control for confounding), 14a (Give characteristics of study participants and information on exposures and potential confounders), and 16a (Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision. Make clear which confounders were adjusted for and why they were included).

Why is this important? As in cohort studies there is no random assignment of the study participants to the different groups to rely on to balance the distribution of confounding factors, it is important to control them either in the design or in the statistical analyses.

Inference 2. Considering the responses given to the items 6 (inference) and 7, is bias between the groups avoided at the beginning of the study?

To avoid competing explanations of the study's results, the groups should be comparable at the beginning of the study. Therefore, it has to be made sure that neither the selection criteria nor the known confounding factors introduced systematic differences between the groups.

Exposure measure

Item 8. Was the exposure explicitly defined?
Item 9. Was the tool used to measure the exposure variable valid?
Item 10. Was the tool used to measure the exposure variable reliable?
Item 11. Was the procedure to measure the exposure the same for all participants?
Inference 3. Could the classification of the participants into exposed or unexposed be biased?

Item 8: Was the exposure explicitly defined?

To answer this item positively (‘Yes’), there has to be a clear and operative definition of the exposure variable. Also if the exposure variable is defined by the measurement tool (e.g. depression as measured by certain instrument) answer this item positively (‘Yes’). If no definition of the exposure is reported or if it is broad and unspecific answer this item negatively (‘No’).

The information assessed by this item should be available if the study was reported following the STROBE statement, as it is required in the item 7: Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable.

Why is this important? It is important to make a clear definition of what exposure is being studied because a single concept might be ambiguous (e.g. intelligence or personality traits) or inaccurate (e.g. diseases with several stages). A clear definition of exposure is important to evaluate the appropriateness of the measurement tools and increase the replicability of the study.

Item 9. Was the tool used to measure the exposure variable valid?

If the tool was validated in a different study (with a different sample) it is important to consider how comparable the two samples (the one used to validate the tool and the one studied) are.

To answer this item positively ('Yes'), the exposure variable should be objective enough to make a validation of the measurement tool unnecessary, or the measurement tool should be validated (for example with other measurements or with an internal consistency score) and an index of its validity should be given. Answer this item as 'Presumably' if the tool used to measure the exposure variable seems valid but no index of its validity is given. If the tool used to measure the exposure variable does not seem valid, or the construct measured by the measuring tool does not match the definition of the exposure variable; then answer this item negatively ('Doubtfully').

The information assessed by this item should be available if the study was reported following the STROBE statement, as it is required in the item 8: For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group.

Why is this important? If there is no evidence or indicator that states that the used instrument is valid it may be difficult to argue that it really measures what it says it measures.

Item 10. Was the tool used to measure the exposure variable reliable?

If the tool was validated in a different study (with a different sample) it is important to consider how comparable the two samples (the one used to validate the tool and the one studied) are.

To answer this item positively ('Yes'), the measurement tool should be either objective enough to make a validation unnecessary, o should have a good inter-rater (if it is hetero-administered) or test-retest reliability (if it is self-administered). Answer this item as 'Presumably' if the tool used to measure the exposure variable seems or is said to be reliable, but no index of its reliability is given. If it is doubtful that the measurement tool is reliable then answer this item negatively ('doubtfully').

The information assessed by this item should be available if the study was reported following the STROBE statement, as it is required in the item 8: For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group.

Why is this important? The reliability of a measure reflects how stable it is. Depending on the characteristics of the measure a different reliability score may be more appropriate than other.

Item.11. Was the procedure to measure the exposure the same for all participants?

To answer this item positively (‘Yes’), the procedure to measure the exposure variable should be the same for all participants. This implies that both the measurement tools should be the same (or equivalent) and should be applied following the same protocol in all cases. Answer this item negatively (‘No’) if the exposure variable is measured in different ways in the different groups. If the procedure by which the exposure variable was measured is not reported then answer this item as ‘Not reported’.

The information assessed by this item should be available if the study was reported following the STROBE statement, as it is required in the item 8: For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group.

Why is this important? Although the label given to the measured construct by different measurement tools might be the same, it is possible that they don’t really measure the same construct. Therefore, if different tools are used, their equivalence has to be proven in order to make sure that they do not introduce systematic differences between the groups. Would there have been the same classification of subjects into exposed and unexposed if the measurement instruments had been applied the other way round? Furthermore, if the measurement instruments are the same or equivalent for all groups but are applied differently in the different groups, it is possible that this systematic difference could be a competing explanation of the results.

Inference 3. Considering the answers given to the items 8 to 11, could the classification of the participants into exposed or unexposed be biased?

In order to be sure that the participants were not misclassified into exposed and unexposed, the construct assessed by the used tool should match the definition of the exposure variable. Additionally, the tool used to measure the exposure variable should be valid and reliable and be applied equally in all groups.

Maintenance of the comparability

Item 12. Were potential confounders that appeared during the follow-up time taken into account in the analyses?
Item 13. Was the length of follow-up similar between the groups?
Item 14. Is there any potential confounder that could have appeared during follow-up that was not taken into account by the authors?
Inference 4. Could the exposure to other factors appearing during follow-up introduce systematic differences between the groups?

Item 12. Were potential confounders that appeared during the follow-up time taken into account in the analyses?

To answer this item positively ('Yes') the authors should have considered the possibility that other confounders could appear during follow-up and have taken measures to control them. For example, if the appearance of confounding factors during the follow-up is recorded and analyzed this item should be answered positively. Answer this item negatively ('No') if the possibility that other confounders could appear during follow-up was not taken into account by the authors.

The information assessed by this item should be available if the study was reported following the STROBE statement, as it is required in the items 7 (Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable), 9 (Describe any efforts to address potential sources of bias), 12a (Describe all statistical methods, including those used to control for confounding), and/or 16a (Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision. Make clear which confounders were adjusted for and why they were included).

Why is this important? As in cohort studies participants are followed up during long periods of time in an uncontrolled environment (real life), it is possible that some elements appear that could affect the outcome or be a competing explanation for the results. As it is not possible to control the environment in observational studies, it is important to assess the confounding variables that may appear during the follow-up period to be able to analyze the scope of its impact on the results.

Item 13. Was the length of follow-up similar between the groups?

To answer this item positively (‘Yes’), the length of follow-up should be the same or similar enough for all groups. Answer this item as ‘No, but controlled’ if the length of follow-up was not similar for all groups but it was taken into account during the analyses. If the length of follow-up was not similar for all groups then answer this item negatively (‘No’).

The information assessed by this item should be available if the study was reported following the STROBE statement, as it is required in the item 14c: Summarize follow-up time (e.g., average and total amount).

Why is this important? If the length of follow-up is different between the groups it may easily be an alternative explanation to the results. For example, in participants with a shorter follow-up time the outcome may not have been developed yet.

Item 14. Is there any potential confounder that could have appeared during follow-up that was not taken into account by the authors?

To answer this item negatively (‘Probably none important’) it should be unlikely that there was any important potential confounder appearing during follow-up that was not taken into account by the authors. Answer this item as ‘Probably’ if, considering the topic under study and the length of follow-up, there probably is a confounding factor that appeared during follow-up and was not taken into account. If there is an already known potential confounder appearing during the follow-up time that was not taken into account by the authors then answer this item positively (‘Yes’).

This item refers to information not taken into account by the authors, so it is not possible that it is reported.

Why is this important? Depending on the topic under study and the length of the follow-up it might be risky to consider that the outcome was not affected by any unknown potential confounder despite all the known confounders have been taken into account.

Inference 4. Considering the responses given to the items 12 to 14, could the exposure to other factors appearing during follow-up introduce systematic differences between the groups?

The comparability of the groups should be maintained along the study in order to avoid distorted results and competing explanations. Therefore, the potential confounders appearing during the follow-up time have to be measured and taken into account during the analyses. If there is any potential confounding factor not considered by the authors, careful reflection about its potential impact on the comparability of the groups is advised.

Outcome measure

If more than one outcome is studied in the systematic review and/or the meta-analysis, then each outcome has to be assessed independently when answering the following items.

Item 15. Was the outcome variable explicitly defined?
Item 16. Was the tool used to assess the outcome variable valid?
Item 17. Was the tool used to assess the outcome variable reliable?
Item 18 (inference). Was the tool used to assess the outcome appropriate?
Item 19. Was the outcome variable assessed in the same way in all groups?
Item 20. Was the outcome variable assessed at the same time for all groups?
Item 21. Was the outcome variable assessed in the same context for all groups?
Item 22 (inference). Could the procedures for measuring the outcome variable introduce systematic differences between the groups?
Item 23. Were the participants successfully blinded to the research question?
Item 24. Were those assessing the outcome successfully blinded to the exposure status of the participants?
Inference 5. Does the measure of the outcome variable reflect the true situation?

Item 15. Was the outcome variable explicitly defined?

To answer this item positively ('Yes'), there should be a clear and operative definition of the outcome variable or it should be defined by the measurement tool (e.g. depression as measured by the BDI). If no definition of the outcome variable is reported or if the definition is broad and unspecific then this item should be answered negatively ('No').

The information assessed by this item should be available if the study was reported following the STROBE statement, as it is required in the items 7: Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable

Why is this important? It is important to make a clear definition of what outcome is being studied because a single concept might be ambiguous or inaccurate. A clear definition of the outcome is important to evaluate the appropriateness of the measurement tool and increase the replicability of the study.

Item 16. Was the tool used to assess the outcome variable valid?

If the tool was validated in a different study (with a different sample) it is important to considerer how comparable the two samples (the one used to validate the tool and the one studied) are.

To answer this item positively ('Yes'), the outcome variable should be objective enough to make a validation of the measurement tool unnecessary, or the measurement tool should be validated (for example with other measurements or with an internal consistency score) and an index of its validity given. Answer this item as 'Presumably' if the tool used to measure the outcome variable seems or is said to be valid, but no index of its validity is given. If the construct measured by the measuring tool does not match with the definition of the outcome variable or the given index of validity is too low.

The information assessed by this item should be available if the study was reported following the STROBE statement, as it is required in the item 8: For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group.

Why is this important? If there is no evidence or indicator that states that the used instrument is valid it is difficult to argue that it really measures what it says it measures.

Item 17. Was the tool used to assess the outcome variable reliable?

If the tool was validated in a different study (with a different sample) it is important to considerer how comparable the two samples (the one used to validate the tool and the one studied) are.

To answer this item positively ('Yes'), the measurement tool should either be objective enough to make a validation unnecessary, or have a good inter-rater (if it is hetero-administered) or test-retest reliability (if it is auto-administered). Answer this item as 'Presumably' if the tool used to measure the outcome variable seems or is said to be reliable, but no index of its reliability is given. If it is doubtful that the measurement tool is reliable answer this item as 'Doubtfully'.

The information assessed by this item should be available if the study was reported following the STROBE statement, as it is required in the item 8: For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group.

Why is this important? The reliability of an assessment tool gives information about how stable its measurement is. Depending on the characteristics of the measure a different reliability score may be more appropriate than other.

Item 18 (inference). Considering the response given to the items 15 to 17, was the tool used to assess the outcome appropriate?

It only should be considered appropriate if the construct measured by the tool matches the definition of the outcome variable and if it is both valid and reliable.

Item 19. Was the outcome variable assessed in the same way in all groups?

To answer this item positively ('Yes') the same measurement tools (or proven equivalents) should have been applied in all groups. Answer this item negatively ('No') if different measurement tools are applied to the different groups.

The information assessed by this item should be available if the study was reported following the STROBE statement, as it is required in the item 8: For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group.

Why is this important? If different measurement tools and procedures are used to assess the outcome for the different groups the comparability of the groups may not be guaranteed. If the measures of one group are more sensible than the ones of another group, bias may arise.

Item 20. Was the outcome variable assessed at the same time for all groups?

To answer this item positively ('Yes') the outcome measures should have been taken at the same time for all groups. Answer this item negatively ('No') if the outcome measures were taken at different times for all groups.

The information assessed by this item should be available if the study was reported following the STROBE statement, as it is required in the item 5: Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection.

Why is this important? If the outcome measures are not taken over the same period of time for all groups they are exposed to several threats: For example, external influences appearing between the measurements of the different groups would potentially affect only the groups pending to assess. On the other hand, if the groups are assessed consecutively the increasing expertise of the assessors could introduce bias and their successful masking would be threatened.

Item 21. Was the outcome variable assessed in the same context for all groups?

To answer this item positively ('Yes') the outcome measures should have been made in the same context (setting, location, etc.) for all groups. Answer this item negatively ('No') if the outcome measures were taken in different contexts for each group.

The information assessed by this item should be available if the study was reported following the STROBE statement, as it is required in the item 5: Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection.

Why is this important? The outcome measure can depend heavily on the context in which it is measured, especially if it is not objective.

Item 22 (inference). Considering the responses given to the items 19 to 21, could the procedures for measuring the outcome variable introduce systematic differences between the groups?

If the way in which the outcome measure is assessed varies systematically between the groups, it could be a competing explanation of the results or a confounding factor. To answer this item the user has to judge if the outcomes were measured similarly enough between the groups, considering the measurement tools and the moment and context of the assessment.

Item 23. Were the participants successfully blinded to the research question?

The fact that the researchers used a procedure to blind the study participants does not necessarily mean that they were unaware of it (participants could not be blind to the research question in spite of being blinded to it). Therefore, the user should judge the possible success of the blinding method carefully.

To answer this item positively (‘Yes’) some procedure should have been successfully used to blind the study participants to the research question. Answer this item negatively (‘No’) if, although necessary, no blinding procedure was reported or its failure is evident. If it is unlikely that the outcomes could be affected by the participants' knowledge of the research question (for example because it is an objective information) then this item should be answered as ‘Not necessary’. If the research question of the analyzed study was not defined at the moment when the outcome was measured, then answer this item as 'Not necessary', too.

The information assessed by this item should be available if the study was reported following the STROBE statement, as it is required in the items 8 (For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group) and/or 9 (Describe any efforts to address potential sources of bias).

Why is this important? Participants, in order to please the researchers, might bias their responses to support the study hypothesis. This is especially important in outcome variables that are subjective or that require a subjective assessment (e.g. the participant may report feeling healthier than he really does).

Item 24. Were those assessing the outcome successfully blinded to the exposure status of the participants?

The fact that the researchers used a procedure to blind the interviewers and data collectors does not necessarily mean that they were unaware of it (they could not be blind to the exposure status of the participants despite being blinded to it). Therefore, the user should judge the possible success of the blinding method carefully.

To answer this item positively (‘Yes’) some procedure has been successfully used to blind those assessing the outcome to the exposure status of the participants. Answer this item negatively (‘No’) if, although necessary, no blinding procedure was reported or its failure is evident. If it is unlikely that the outcomes could be affected by the assessors' knowledge of the participant's condition (for instance, when the outcome is objective) then answer this item as ‘Not necessary’.

The information assessed by this item should be available if the study was reported following the STROBE statement, as it is required in the items 8 (For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group) and/or 9 (Describe any efforts to address potential sources of bias).

Why is this important? If those collecting the data are aware of the condition of the participants there’s the risk that they treat them in a different way, making them susceptible to conscious or unconsciously try to confirm the study hypothesis.

Inference 5. Considering the responses given to the items 18 (inference), 22 (inference), and 23 to 24, does the measure of the outcome variable reflect the true situation?

To be sure that the measures of the outcome variable reflect the true situation, the tool used should be appropriate and bias should be avoided in the measuring procedure and by successfully blinding (when necessary) the participants and those measuring the outcomes.

Attrition

Item 25. Were drop out rates similar in all groups?
Item 26. Were reasons for dropping out similar in all groups?
Inference 6. Could incomplete information introduce systematic differences between groups?

Item 25: Were drop out rates similar in all groups?

It is important to differentiate drop outs from missing data, where only certain information is missing.

To answer this item positively (‘Yes’) the percentage of participants dropping out or lost to follow-up should be similar in all groups. Answer this item negatively (‘No’) if the percentage of participants dropping out or lost to follow-up is not similar in all groups. If the drop out rates in each group are not reported and cannot be calculated then answer this item as ‘Not reported’.

The information assessed by this item should be available if the study was reported following the STROBE statement, as it is required in the item 13a: Report numbers of individuals at each stage of study—e.g. numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analyzed.

Why is this important? There should not be drop outs rates that are bigger in one group than in another. Drop out rates may be higher or lower, but if they differ significantly between the groups it is probable that the participants who drop out are not at random, but that there is a variable (or group of variables) affecting differently the different groups. If this is the case, it is important to control this influence, which may confound the results, in the statistical analysis.

Item 26. Were reasons for dropping out similar in all groups?

To answer this item positively (‘Yes’) the reasons for dropping out should be similar between the groups. Answer this item negatively (‘No’) if the reasons for dropping out were different between the groups. If the reasons for dropping out are not reported then answer this item as ‘Not reported’.

The information assessed by this item should be available if the study was reported following the STROBE statement, as it is required in the item 13b: Give reasons for non-participation at each stage.

Why is this important? While two groups may have similar drop out rates, what is more important is to know if the reasons for the loss of participants are different between them, because if they differ systematically, this could lead to bias.

Inference 6. Considering the responses given to the items 25 and 26, could incomplete information introduce systematic differences between groups?

In this item the user has to judge if the comparability of the groups could be affected by selective drop outs. While a low drop out rate is desired, what is really important is to make sure that the underlying reasons are similar, as they might point out systematic differences between the groups. If the reasons for the participants dropping out of the study are not reported, differences in the drop out rates may be an indicator of systematic differences between the groups. In the case that the drop out rates are not given for the different groups, a low drop out is necessary to have enough confidence that the comparability of the groups remains at the end of the study.

Statistical analyses

Inference 7. Do the results of the statistical analysis reflect the true situation?

To judge this aspect the user has to consider if the use of inappropriate statistical procedures could have led to results that do not reflect the true situation. Among others, the user has to consider if the treatment of incomplete data corresponds to the study question, if adjustments were made for multiple comparisons, and the use of any methods used to examine subgroups and interactions.

The information assessed by this item should be available if the study was reported following the STROBE statement, as it is required in the items 12 a-e.

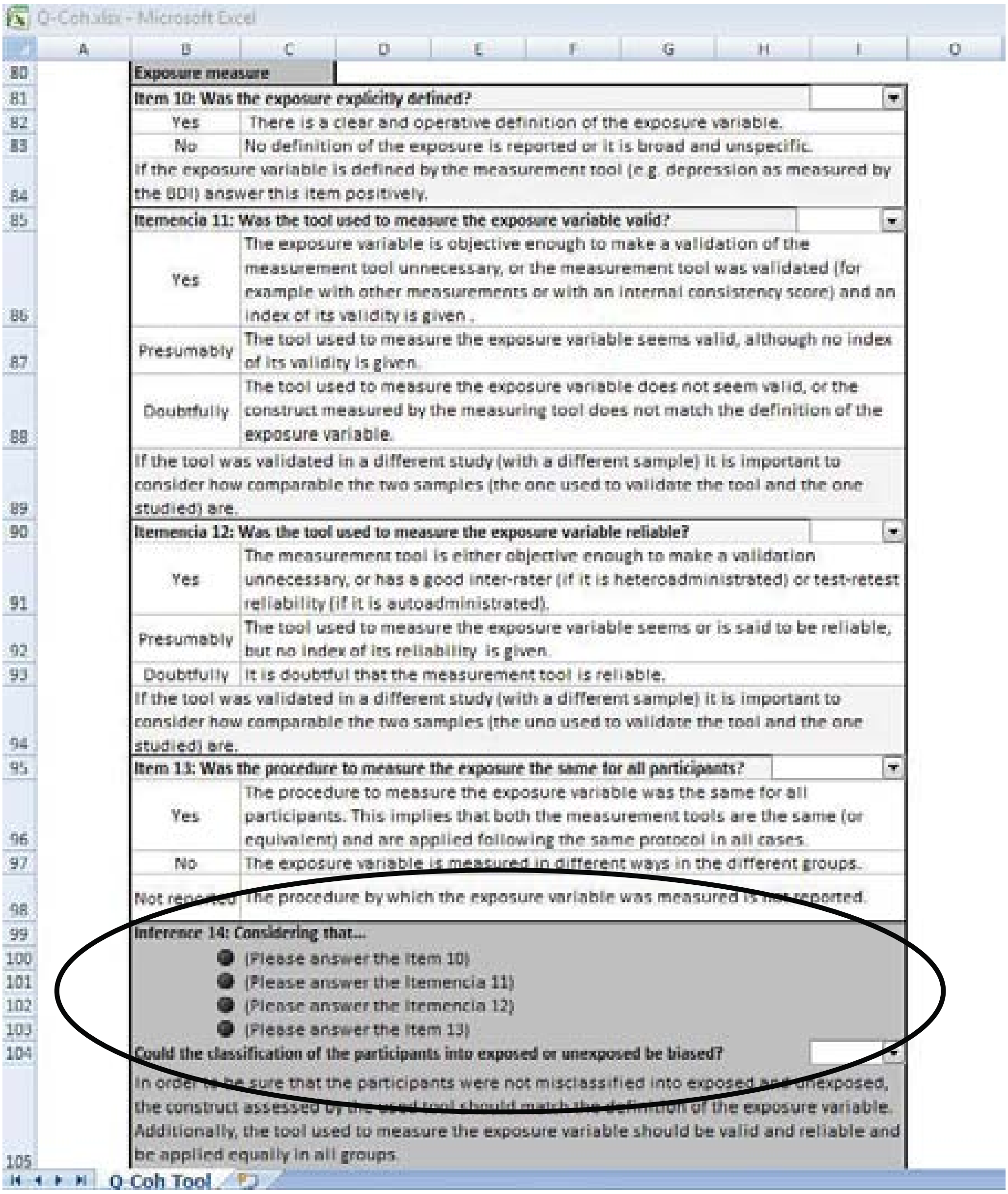
Why is this important? Although if the data collected is of good quality, if it is elaborated using wrong or inappropriate procedures, then they will not reflect the true situation.

Overall assessment of the study’s quality

Considering the judgments made at the inferences 1 to 7, What overall quality does this study have?

Annex 5: Screenshots of the Q-Coh (Excel spreadsheet)

The following screenshots of the Q-Coh show how the use of an Excel spreadsheet allows a more user-friendly interface and avoids that the user has to remember his previous answers or look them up when he has to make any judgments summarizing the previous items.



Exposure measure

Item 10: Was the exposure explicitly defined?

Yes	There is a clear and operative definition of the exposure variable.
No	No definition of the exposure is reported or it is broad and unspecific.

If the exposure variable is defined by the measurement tool (e.g. depression as measured by the BDI) answer this item positively.

Itemencia 11: Was the tool used to measure the exposure variable valid?

Yes	The exposure variable is objective enough to make a validation of the measurement tool unnecessary, or the measurement tool was validated (for example with other measurements or with an internal consistency score) and an index of its validity is given .
Presumably	The tool used to measure the exposure variable seems valid, although no index of its validity is given.
Doubtfully	The tool used to measure the exposure variable does not seem valid, or the construct measured by the measuring tool does not match the definition of the exposure variable.

If the tool was validated in a different study (with a different sample) it is important to consider how comparable the two samples (the one used to validate the tool and the one studied) are.

Itemencia 12: Was the tool used to measure the exposure variable reliable?

Yes	The measurement tool is either objective enough to make a validation unnecessary, or has a good inter-rater (if it is heteroadministrated) or test-retest reliability (if it is autoadministrated).
Presumably	The tool used to measure the exposure variable seems or is said to be reliable, but no index of its reliability is given.
Doubtfully	It is doubtful that the measurement tool is reliable.

If the tool was validated in a different study (with a different sample) it is important to consider how comparable the two samples (the uno used to validate the tool and the one studied) are.

Item 13: Was the procedure to measure the exposure the same for all participants?

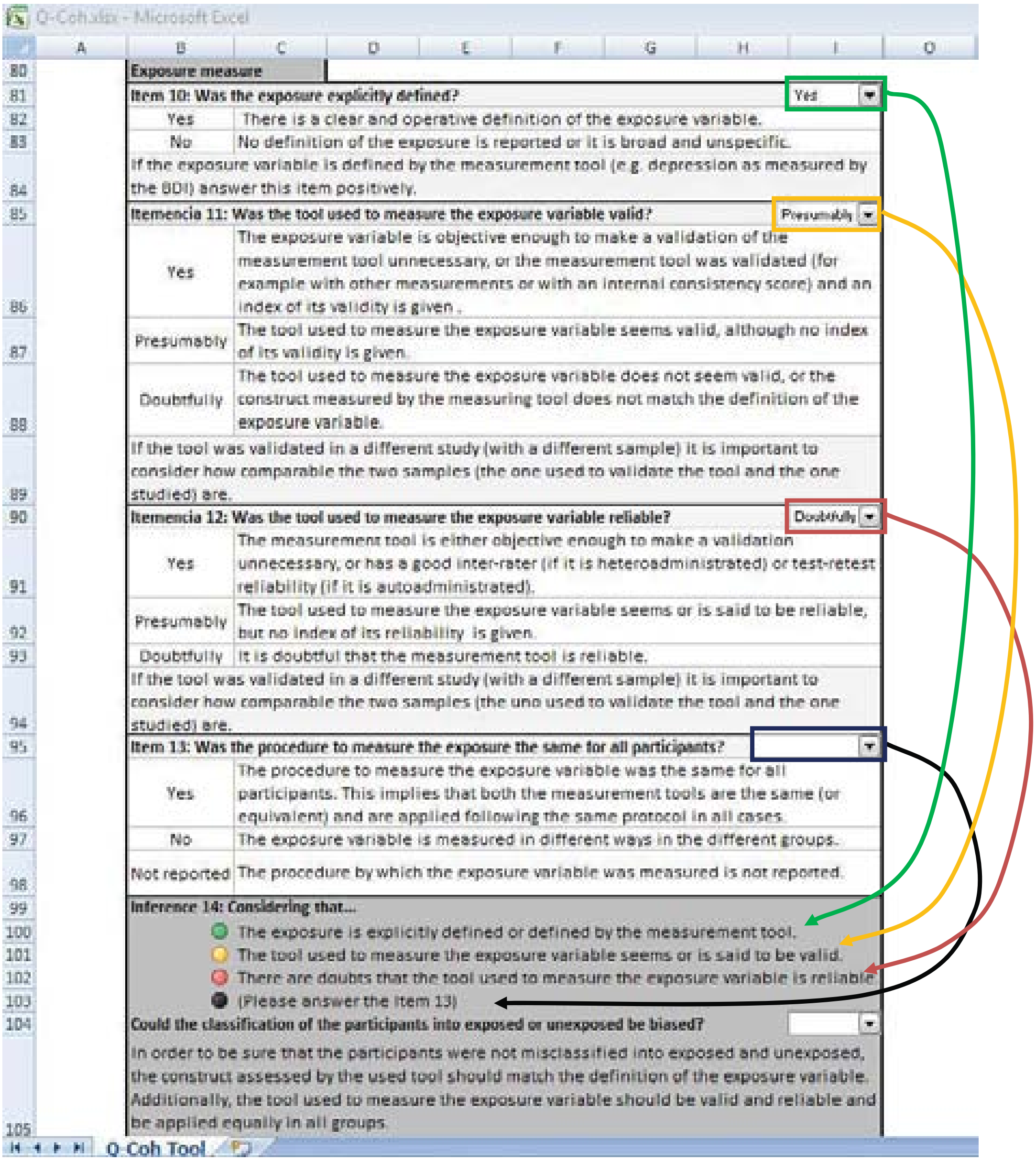
Yes	The procedure to measure the exposure variable was the same for all participants. This implies that both the measurement tools are the same (or equivalent) and are applied following the same protocol in all cases.
No	The exposure variable is measured in different ways in the different groups.
Not reported	The procedure by which the exposure variable was measured is not reported.

Inference 14: Considering that...

- ● (Please answer the item 10)
- ● (Please answer the Itemencia 11)
- ● (Please answer the Itemencia 12)
- ● (Please answer the item 13)

Could the classification of the participants into exposed or unexposed be biased?

In order to be sure that the participants were not misclassified into exposed and unexposed, the construct assessed by the used tool should match the definition of the exposure variable. Additionally, the tool used to measure the exposure variable should be valid and reliable and be applied equally in all groups.



Exposure measure

Item 10: Was the exposure explicitly defined? Yes

Yes	There is a clear and operative definition of the exposure variable.
No	No definition of the exposure is reported or it is broad and unspecific.

If the exposure variable is defined by the measurement tool (e.g. depression as measured by the BDI) answer this item positively.

Itemencia 11: Was the tool used to measure the exposure variable valid? Presumably

Yes	The exposure variable is objective enough to make a validation of the measurement tool unnecessary, or the measurement tool was validated (for example with other measurements or with an internal consistency score) and an index of its validity is given .
Presumably	The tool used to measure the exposure variable seems valid, although no index of its validity is given.
Doubtfully	The tool used to measure the exposure variable does not seem valid, or the construct measured by the measuring tool does not match the definition of the exposure variable.

If the tool was validated in a different study (with a different sample) it is important to consider how comparable the two samples (the one used to validate the tool and the one studied) are.

Itemencia 12: Was the tool used to measure the exposure variable reliable? Doubtfully

Yes	The measurement tool is either objective enough to make a validation unnecessary, or has a good inter-rater (if it is heteroadministrated) or test-retest reliability (if it is autoadministrated).
Presumably	The tool used to measure the exposure variable seems or is said to be reliable, but no index of its reliability is given.
Doubtfully	It is doubtful that the measurement tool is reliable.

If the tool was validated in a different study (with a different sample) it is important to consider how comparable the two samples (the uno used to validate the tool and the one studied) are.

Item 13: Was the procedure to measure the exposure the same for all participants?

Yes	The procedure to measure the exposure variable was the same for all participants. This implies that both the measurement tools are the same (or equivalent) and are applied following the same protocol in all cases.
No	The exposure variable is measured in different ways in the different groups.
Not reported	The procedure by which the exposure variable was measured is not reported.

Inference 14: Considering that...

🟢 The exposure is explicitly defined or defined by the measurement tool.
🟡 The tool used to measure the exposure variable seems or is said to be valid.
🔴 There are doubts that the tool used to measure the exposure variable is reliable
⚫ (Please answer the item 13)

Could the classification of the participants into exposed or unexposed be biased?

In order to be sure that the participants were not misclassified into exposed and unexposed, the construct assessed by the used tool should match the definition of the exposure variable. Additionally, the tool used to measure the exposure variable should be valid and reliable and be applied equally in all groups.

Annex 6: Preview of the Q-Coh (Appendix of Article 3)

Preview of the Q-Coh

Design of the study	
Item A: Is there a comparison between at least two groups to assess the effect/ association of an exposure and an outcome?	
Yes	The effect of the exposure and the outcome was estimated for at least two groups.
No	The effect of the exposure and the outcome was NOT estimated for at least two groups.
This means that there have to be at least two measures of the effect/ association of an exposure and an outcome that can be compared. If there is only one measure of association, we could only say how prone those exposed are to have the outcome. If two measures of association are compared, we can say how much more prone those exposed are to have the outcome than those not exposed (relative risk). So, to be able to compute an effect size, a relationship between at least two groups is needed.	
Item B: Are the groups defined by the exposure variable?	
Yes	The groups that are compared are defined by their different value in the exposure variable.
No	The groups that are compared are defined by other variables that are NOT the exposure variable.
This means that the different groups are formed by aggregating subjects exposed and unexposed (or with different degrees of exposure). Subjects may be selected from different populations (each defined by being exposed or not) or from the same population (defined by other selection criteria) and then classified into exposed or unexposed.	
Item C: Has any of the participants the outcome of interest on onset?	
No	None of the participants have the outcome of interest at the time the exposure is measured for the first time.
Yes	At least some of the participants have the outcome of interest at the time the exposure is measured for the first time.
If the outcome is defined by a change in the value of a variable (for example and increase or decrease in perceived anxiety), then this item should be answered 'No'.	
Item D: Do investigators handle who is exposed or not?	
No	Researchers have no control over who is exposed or not (no assignment to exposure).
Yes	Researchers assign the participants to the exposed or unexposed groups.
Item E: Is information about the exposure and the outcome of interest registered concurrently?	
No	The measures of exposure and outcome have been registered at different time points.
Yes	There is no time lapse between the measurement of the exposure and the outcome.
Inference 0: Considering Items A to E…	
Is the tool suitable for this study?	**Probably | Unlikely**
When judging this aspect the user has to consider the answers given to the items A to E. These items screen for the characteristics that define a cohort study as understood here, so if the items A and B are not answered positively ('Yes') or if items C, D or E have not been answered negatively ('No'), then the study has a design not contemplated by this tool and therefore should be assessed with another quality assessment tool. This is so regardless if the study is labeled as a 'cohort' or 'prospective' study by either the authors or the database classification.	

Representativeness

Item 1: Have the study participants been selected using a randomized sampling procedure?

Yes	The study participants have been selected at random from a complete list of the target population or, if they are randomly selected from natural clusters of subjects (*e.g.* hospitals, schools, cities, etc.), these also should have been selected at random among all clusters that form the target population.
No	The study participants or the clusters from where they are selected have been chosen using a non-randomized procedure (*e.g.* by convenience). Therefore, not all subjects of the target population had a chance of being selected.

It is important that the randomized sampling procedure guarantees that all the subjects of the target population have a chance of being selected. Attention should be paid for not getting confused by quasi-randomized sampling procedures.

Item 2: Is the similarity between the selected group of subjects and the target population justified by the authors?

Yes, empirically	The similarity of the selected sample with the target population is empirically stated (*e.g.* comparing them on demographic, clinical or social characteristics).
Yes, verbally	The authors discuss the sampling procedure and justify the representativeness of the selected group of subjects, but without any empirical comparison.
No	The similarity between the selected groups of subjects and the target population is not justified or is stated as a limitation of the study.

If the sampling procedure was not completely randomized, it is important that the authors justify that the selected group of subjects is similar to the target population despite it has not been selected using a randomized procedure.

Item 3: Is there a predominant reason for refusing to participate at the beginning of the study?

No / Irrelevant	Participants and those refusing to enter the study are similar or the number of subjects who refuse to participate is small enough to be considered irrelevant.
Yes	Non-participants have something in common or give similar reasons for refusing to participate in the study.
Not reported	No information is reported about the quantity or the reasons and characteristics of those refusing to participate in the study.

It is important to not confound those who refuse to enter the study with participants who entered the study and dropped out along the follow-up or who refuse to participate when the outcome is assessed.

Inference 1: Considering Items 1 to 3…

Was the sample representative of the target population? **Probably | Unlikely**

For the sample to be representative of the target population, the selected group of subjects should have been selected using a randomized sampling procedure that guarantees that all the subjects of the target population had a chance of being selected, and there should be no predominant reason for refusing entering the study. If the sampling procedure was not completely randomized, then the similarity between the sample and the target population should be justified.

Comparability of the groups

Item 4: Were the inclusion and exclusion criteria explicitly defined for all groups?

Yes	There is a clear and operative definition of the inclusion and exclusion criteria for all groups.
No	The inclusion and exclusion criteria are not reported or are broad and unspecific.

This refers to the criteria each subject has to fulfill in order to be suitable for the study.

Item 5: Were the same inclusion and/or exclusion criteria applied equally to all groups?

Yes	Those sampled individuals who are willing to participate should have been applied the same inclusion and exclusion criteria (except those referring to the exposure variable) in the same way.
No	The selection criteria were not the same or were applied differently in each group (e.g. different screening tools)
Not reported	No selection criteria are reported or if it was not reported how they were applied.

Item 6: Considering Items 4 and 5…

Could differences in the selection criteria introduce systematic differences between the groups (other than exposure)? **Probably | Unlikely**

In order to answer this inference as 'Unlikely' the selection criteria should be explicitly defined and be applied equally to all groups. If the selection criteria are not explicitly defined, if they are different for the different groups, or if they are not equally applied between them, then it has to be carefully considered in how far each of these threats to the comparability may introduce systematic differences between the groups.

Item 7: Were known confounding factors accounted for in the design or in the analysis?

Yes	All the important confounding variables have been accounted for in the design or in the statistical analyses.
Partially	Some, but not all the important confounding variables have to be accounted for in the design or in the statistical analyses.
No	None of the important confounding variables have been accounted for in the design or in the statistical analysis.

Considering that this tool is to be used in the context of systematic reviews and meta-analyses, the most important confounding variables should have been defined previously.

There are several ways of taking a confounding factor into account in the design: matching, stratification or restriction. The objective of these three procedures is to make the groups comparable on known confounding factors, either assuring a similar distribution of them (matching and stratification) or maintaining them constant (restriction).

Although if a confounding variable has not been accounted for in the design, it can still be controlled during the data analysis if it has been measured, for example using stratification, adjustment, propensity scores or in regression models.

Inference 2: Considering Items 6 and 7…

Is bias between the groups avoided at the beginning of the study? **Probably | Unlikely**

To avoid competing explanations of the study's results, the groups should be comparable at the beginning of the study. Therefore, it has to be made sure that either the selection criteria, nor the known and unknown confounding factors introduced systematic differences between the groups

Exposure measure	
Item 8: Was the exposure explicitly defined?	
Yes	There is a clear and operative definition of the exposure variable.
No	No definition of the exposure is reported or it is broad and unspecific.
If the exposure variable is defined by the measurement tool (e.g. depression as measured by the BDI) answer this item positively.	
Item 9: Was the tool used to measure the exposure variable valid?	
Yes	The exposure variable is objective enough to make a validation of the measurement tool unnecessary, or the measurement tool was validated (for example with other measurements or with an internal consistency score) and an index of its validity is given .
Presumably	The tool used to measure the exposure variable seems valid, although no index of its validity is given.
Doubtfully	The tool used to measure the exposure variable does not seem valid, or the construct measured by the measuring tool does not match the definition of the exposure variable.
If the tool was validated in a different study (with a different sample) it is important to consider how comparable the two samples (the one used to validate the tool and the one studied) are.	
Item 10: Was the tool used to measure the exposure variable reliable?	
Yes	The measurement tool is either objective enough to make a validation unnecessary, or has a good inter-rater (if it is heteroadministrated) or test-retest reliability (if it is autoadministrated).
Presumably	The tool used to measure the exposure variable seems or is said to be reliable, but no index of its reliability is given.
Doubtfully	It is doubtful that the measurement tool is reliable.
If the tool was validated in a different study (with a different sample) it is important to consider how comparable the two samples (the one used to validate the tool and the one studied) are.	
Item 11: Was the procedure to measure the exposure the same for all participants?	
Yes	The procedure to measure the exposure variable was the same for all participants. This implies that both the measurement tools are the same (or equivalent) and are applied following the same protocol in all cases.
No	The exposure variable is measured in different ways in the different groups.
Not reported	The procedure by which the exposure variable was measured is not reported.
Inference 3: Considering Items 8 to 11… **Could the classification of the participants into exposed or unexposed be biased?**	**Probably | Unlikely**
In order to be sure that the participants were not misclassified into exposed and unexposed, the construct assessed by the used tool should match the definition of the exposure variable. Additionally, the tool used to measure the exposure variable should be valid and reliable and be applied equally in all groups.	

Maintenance of the comparability

Item 12: Were potential confounders that appeared during the follow-up time taken into account in the analyses?

Yes	The authors have considered the possibility that other confounders could appear during follow-up and have taken measures to control them.
No	The possibility that other confounders could appear during follow-up was not taken into account by the authors.

As an example, if the appearance of confounding factors during the follow-up is recorded and analyzed this item should be answered as 'Yes'.

Item 13: Was the length of follow-up similar between the groups?

Yes	The length of follow-up was the same or similar enough for all groups.
No, but controlled	The length of follow-up was not similar for all groups but it was taken into account during the analyses.
No	The length of follow-up was not similar for all groups.

Item 14: Is there any potential confounder that could have appeared during follow-up that was not taken into account by the authors?

Probably none important	It is unlikely that there was any important potential confounder appearing during follow-up that was not taken into account by the authors.
Probably	Considering the topic under study and the length of follow-up, there was probably a confounding factor that appeared and during follow-up and was not taken into account.
Yes	There is a known potential confounder appearing during the follow-up time that was not taken into account by the authors.

Inference 4: Considering Items 12 to 14…

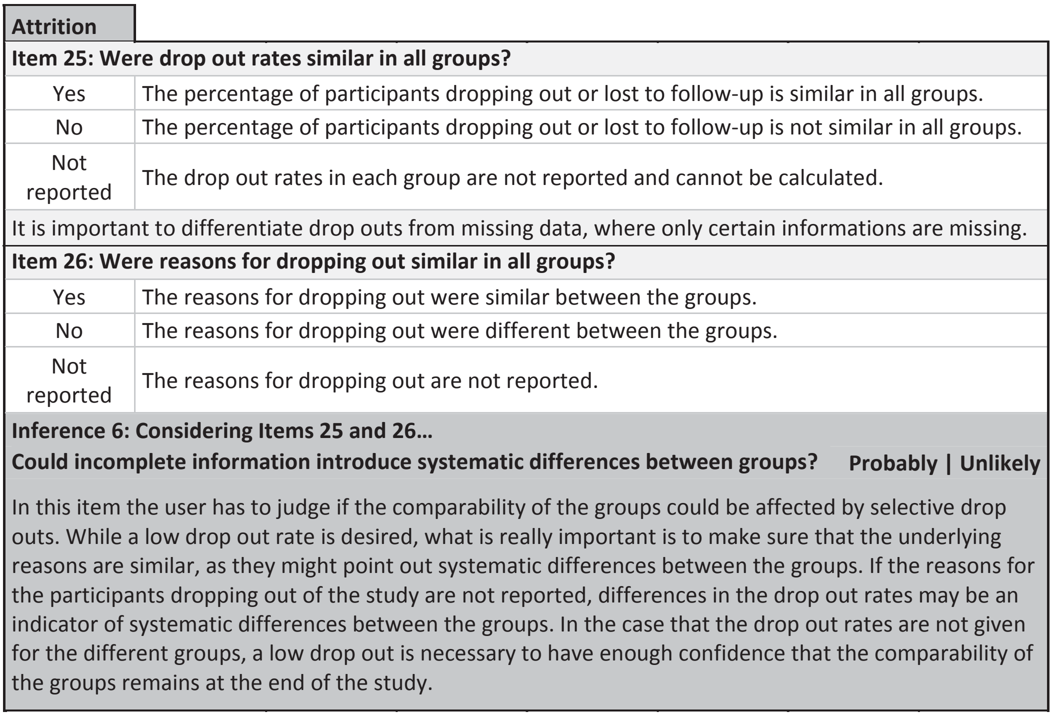
Could the exposure to other factors appearing during follow-up introduce systematic differences between the groups? **Probably | Unlikely**

The comparability of the groups has to be maintained along the study in order to avoid distorted results and competing explanations. Therefore, the potential confounders appearing during the follow-up time have to be measured and taken into account during the analyses. If there is any potential confounding factor not considered by the authors, careful reflection about its potential impact on the comparability of the groups is advised.

Outcome measure

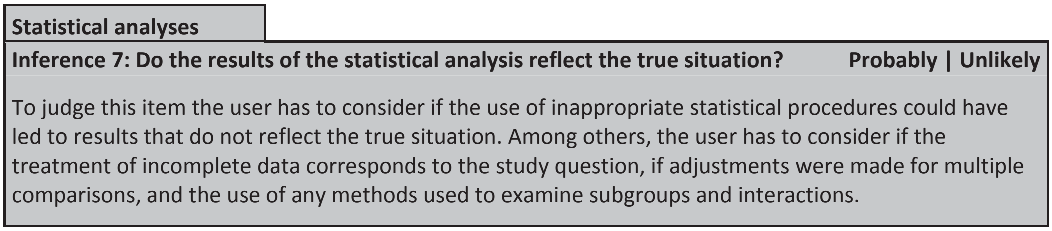
If more than one outcome is studied in the systematic review and/or the meta-analysis, then each outcome has to be assessed independently when answering the following items.	
Item 15: Was the outcome variable explicitly defined?	
Yes	There is a clear and operative definition of the outcome variable or it is defined by the measurement tool (e.g. depression as measured by the BDI).
No	No definition of the outcome variable is reported or if it is broad and unspecific.
Item 16: Was the tool used to assess the outcome variable valid?	
Yes	The outcome variable is objective enough to make a validation of the measurement tool unnecessary, or the measurement tool was validated (for example with other measurements or with an internal consistency score) and an index of its validity is given.
Presumably	The tool used to measure the outcome variable seems or is said to be valid, but no index of its validity is given.
Doubtfully	The construct measured by the measuring tool does not match with the definition of the outcome variable, or the given index of validity is too low.
If the tool was validated in a different study (with a different sample) it is important to considerer how comparable the two samples (the one used to validate the tool and the one studied) are.	
Item 17: Was the tool used to assess the outcome variable reliable?	
Yes	The measurement tool is either objective enough to make a validation unnecessary, or has a good inter-rater (if it is heteroadministrated) or test-retest reliability (if it is autoadministrated).
Presumably	The tool used to measure the outcome variable seems or is said to be reliable, but no index of its reliability is given.
Doubtfully	It is doubtful that the measurement tool is reliable.
If the tool was validated in a different study (with a different sample) it is important to considerer how comparable the two samples (the one used to validate the tool and the one studied) are.	
Item 18: Considering Items 15 to 17… **Was the tool used to assess the outcome appropriate?**	**Probably | Unlikely**
If the construct measured by the tool matches the definition of the outcome variable and if it is valid and reliable, then it should only be considered appropriate.	

<table>
<tr><td colspan="2"><b>Item 19: Was the outcome variable assessed in the same way in all groups?</b></td></tr>
<tr><td>Yes</td><td>The same measurement tools (or proven equivalents) have been applied in all groups.</td></tr>
<tr><td>No</td><td>Different measurement tools are applied to the different groups.</td></tr>
<tr><td colspan="2"><b>Item 20: Was the outcome variable assessed at the same time for all groups?</b></td></tr>
<tr><td>Yes</td><td>The outcome measures were taken at the same time for all groups.</td></tr>
<tr><td>No</td><td>The outcome measures were taken at different times for all groups.</td></tr>
<tr><td colspan="2"><b>Item 21: Was the outcome variable assessed in the same context for all groups?</b></td></tr>
<tr><td>Yes</td><td>The outcome measures were made in the same context (setting, location, etc.) for all groups.</td></tr>
<tr><td>No</td><td>The outcome measures were taken in different contexts for all groups.</td></tr>
<tr><td colspan="2"><b>Item 22: Considering Items 19 to 21…</b>
<b>Could the procedures for measuring the outcome variable introduce systematic differences between the groups?</b> <b>Probably | Unlikely</b>
To answer this item the user has to judge if the outcomes were measured similarly enough between the groups, considering the measurement tools and the moment and context of the assessment.</td></tr>
<tr><td colspan="2"><b>Item 23: Were the participants successfully blinded to the research question?</b></td></tr>
<tr><td>Yes</td><td>Some procedure has been successfully used to blind the study participants to the research question.</td></tr>
<tr><td>No</td><td>Although necessary, no blinding procedure was reported or its failure is evident.</td></tr>
<tr><td>Not necessary</td><td>It is unlikely that the outcomes could be affected by the participants' knowledge of the research question (for example because it is an objective information).</td></tr>
<tr><td colspan="2">If the research question of the analyzed study was not defined at the moment when the outcome was measured, then answer this item as 'Not necessary'.</td></tr>
<tr><td colspan="2"><b>Item 24: Were those assessing the outcome successfully blinded to the exposure status of the participants?</b></td></tr>
<tr><td>Yes</td><td>Some procedure has been successfully used to blind those assessing the outcome to the exposure status of the participants.</td></tr>
<tr><td>No</td><td>Although necessary, no blinding procedure was reported or its failure is evident.</td></tr>
<tr><td>Not necessary</td><td>It is unlikely that the outcomes could be affected by the assessors' knowledge of the participant's condition (for instance, when the outcome is objective).</td></tr>
<tr><td colspan="2"><b>Inference 5: Considering Items 18, and 22 to 24…</b>
<b>Does the measure of the outcome variable reflect the true situation?</b> <b>Probably | Unlikely</b>
To be sure that the measures of the outcome variable reflect the true situation, the tool used should be appropriate and bias should be avoided in the measuring procedure and by successfully blinding (when necessary) the participants and those measuring the outcomes.</td></tr>
</table>



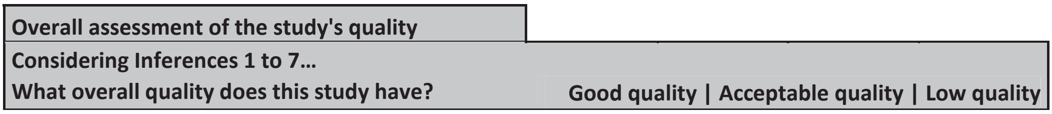
Attrition

Item 25: Were drop out rates similar in all groups?	
Yes	The percentage of participants dropping out or lost to follow-up is similar in all groups.
No	The percentage of participants dropping out or lost to follow-up is not similar in all groups.
Not reported	The drop out rates in each group are not reported and cannot be calculated.
It is important to differentiate drop outs from missing data, where only certain informations are missing.	
Item 26: Were reasons for dropping out similar in all groups?	
Yes	The reasons for dropping out were similar between the groups.
No	The reasons for dropping out were different between the groups.
Not reported	The reasons for dropping out are not reported.
Inference 6: Considering Items 25 and 26… **Could incomplete information introduce systematic differences between groups?**	**Probably | Unlikely**
In this item the user has to judge if the comparability of the groups could be affected by selective drop outs. While a low drop out rate is desired, what is really important is to make sure that the underlying reasons are similar, as they might point out systematic differences between the groups. If the reasons for the participants dropping out of the study are not reported, differences in the drop out rates may be an indicator of systematic differences between the groups. In the case that the drop out rates are not given for the different groups, a low drop out is necessary to have enough confidence that the comparability of the groups remains at the end of the study.	



Statistical analyses

Inference 7: Do the results of the statistical analysis reflect the true situation?	**Probably | Unlikely**
To judge this item the user has to consider if the use of inappropriate statistical procedures could have led to results that do not reflect the true situation. Among others, the user has to consider if the treatment of incomplete data corresponds to the study question, if adjustments were made for multiple comparisons, and the use of any methods used to examine subgroups and interactions.	



Overall assessment of the study's quality

Considering Inferences 1 to 7… **What overall quality does this study have?**	**Good quality | Acceptable quality | Low quality**

Notes: This is an adaptation of the original tool, which is a Microsoft Excel spreadsheet that allows recording the responses into a database and reminds the answers made to the previous items that have to be considered for answering other and making inferences. We strongly recommend using the Excel version of the Q-Coh Tool, which can be requested to the authors.