

## INVITED COMMENTARY

# Quality appraisal as part of the systematic review: a review of current methods

**C.Littlewood, K.Chance-Larsen, S.M. McLean**

*Faculty of Health and Wellbeing, Sheffield Hallam University, Sheffield, UK, S10 2BP*  
*Email: [c.littlewood@shu.ac.uk](mailto:c.littlewood@shu.ac.uk)*

---

## Abstract

Systematic reviews frequently underpin national and international practice guidelines. Different approaches to the systematic review process, in particular quality appraisal, have been advocated. This paper discusses these approaches and highlights possible limitations which might impact upon the validity of the conclusions drawn. Practical alternatives are offered upon which systematic reviews may be appraised and conducted.

*Keywords:* systematic review, best evidence synthesis, research methods, quality appraisal.

---

## Introduction

Over recent years there has been a dramatic rise in the volume of published literature which makes it almost impossible for relevant stakeholders, including patients, clinicians, researchers and commissioners, to keep abreast of developments in physiotherapy and related fields (Carroll et al 2008). The systematic review is one approach that has been designed to address this difficulty ([Assendelft et al 1995](#)) with the aim of synthesising the findings of primary studies on a single topic in order to assess the overall clinical impact and relevance of that body of literature. This popular research method utilises systematic and transparent means to identify, select, quality appraise and synthesise research and frequently underpins national and international practice guidelines (Hettinga et al 2008, Higgins and Green 2009).

To facilitate the process of quality appraisal and synthesis many systematic reviews in the field of physiotherapy have adopted rating scales to assign a numerical value to the quality of a research study ([Barr et al 2009](#); [Paratz and Stockton 2009](#); [Tang et al 2010](#)). There are many published rating scales available which tend to take the form of a checklist comprised of criteria, e.g. randomisation, level of blinding, which are thought to be important factors in determining the degree of potential bias in the research which in turn would affect the internal and/ or external validity of the study (Maher et al 2003, van Tulder et al 2003). The checklist is used as the basis of a quality appraisal and a value is assigned depending upon the number of criteria met or not met. The study is subsequently judged as being of high, moderate or low quality which is

reflected in the final synthesis where high quality studies carry greater weight. Despite the likelihood of methodological flaws and hence potentially unreliable evidence, low quality studies are usually still included in the synthesis which might influence the overall conclusions drawn in an unpredictable way (Hettinga et al 2008).

## **Discussion**

The validity of different systematic review methods, particularly quality appraisal using numerical rating scales and the inclusion of low quality studies, has been raised in the literature (Hettinga et al 2008, Littlewood and May 2007, Slavin 1995, van der Velde et al 2007). It has been recognised that different approaches to quality appraisal in systematic reviews may yield different results. These results may also be at odds with large randomised controlled trials (RCT's) with the same focus which are regarded as the 'gold' standard of evaluative research by many (LeLorier et al 1997, van der Velde et al 2007).

Firstly, considering the impact of low quality primary studies on the conclusion drawn by a systematic review, it is suggested that inadequate attention to key design features including sample size, allocation and level of blinding (Kjaergard et al 2001, Moore et al 1998) during the planning and execution of research studies may serve to create an over or under exaggeration of diagnostic accuracy or treatment efficacy (Hettinga et al 2008). Hence the inclusion of low quality studies when synthesising data may give rise to inaccurate conclusions. It has been put forward that summing these studies through meta-analysis or a qualitative synthesis may guard against this effect; however this assumption should be challenged because synthesis cannot take into account the direction or extent of bias in individual studies (Kjaergard et al 2001, LeLorier et al 1997, Slavin 1995). Thus, it is perhaps not surprising that many systematic reviews which include low quality studies within their synthesis identify conflicting findings between studies or arrive at unclear conclusions (Ho et al 2009, May et al 2006, van Trijffell et al 2005).

As well as considering the impact of inclusion of low quality studies, it has also been suggested that analysis by quality score neither adjusts nor removes the bias of studies (NHMRC 1999). Hence even studies regarded as high quality determined by meeting a certain number of criteria from a numerical checklist may have key design flaws which render the study scientifically inadmissible. Examples of this exist in published systematic reviews, e.g. Tang et al (2010), where RCT's are regarded as high quality even in the absence of key design features such as concealed allocation, assessor blinding or intention to treat analysis. Because systematic reviews of this nature take into account the quality of the studies when synthesising results and drawing conclusions the arbitrary cut-off point encouraged by the use of checklists might have a significant impact upon the validity of the conclusions drawn.

To address these concerns alternative methods have been proposed. The best-evidence synthesis approach to the systematic review follows the main conventions of other approaches to the systematic review process, but does not assign relative quality values to included studies based upon checklists (Carroll et al 2008, Slavin 1995). Instead, the studies that meet the pre-defined inclusion criteria for the review are appraised by the systematic review team based upon pre-defined guidelines and the studies are judged as being scientifically admissible or not. This process relies on the skill and expertise of the review team, which in itself might be a source of

potential bias, and where a study is not regarded as scientifically admissible it is rejected from the review process.

A study by van der Velde et al (2007) sought to evaluate the impact of different approaches to the systematic review process upon the conclusions of the relevant review. The best evidence-synthesis approach and the approach based upon The Cochrane Back Review Group Guidelines were adopted. The authors, including advocates of both approaches, recognised the methodological shortcomings of the different approaches. With respect to the Cochrane approach, van der Velde et al (2007) suggested that most limitations associated with this approach were related to the use of a checklist to appraise quality. They recognised that only methodological weaknesses included in the Cochrane checklist were evaluated and other factors, e.g. validity of outcome measures used, were overlooked. They also recognised that the use of a non validated cut-off point to assign a quality rating to the included studies, as discussed above, in association with the inclusion of methodologically weak studies was likely to introduce bias into the results of the review. In contrast to this, these authors suggested that the weaknesses associated with the best-evidence synthesis approach lay with a less structured approach to quality appraisal which was reliant upon the composition and competence of the systematic review team. This paper concludes by recognising that different approaches to the systematic review process may result in different outcomes which consumers of the literature need to be aware of when interpreting the conclusions of reviews. Other work, e.g. Hettinga et al (2008) have also identified that different approaches to the systematic review, including when low quality studies are maintained in the review, produce conflicting results.

In summary, this paper has identified three main issues that should be recognised when conducting and appraising systematic reviews. Firstly, including low quality studies with fatal flaws might introduce significant bias into the findings of systematic reviews. Secondly, quality appraisal using numerical checklists does not guard against bias and when employed in isolation might not be a useful way of assigning relative quality. Thirdly, a review process which is reliant upon the skill and composition of a systematic review team making qualitative judgments may not be regarded as a robust and transparent process.

## **Implications for practice**

Despite the limitations we recognise that the systematic review remains superior to the narrative review process which does not include mechanisms to minimise bias. However, the above discussion offers insight into features of systematic review design that we should be mindful of when critically appraising and making decisions about the trustworthiness of the findings to influence clinical decision making. Clearly there is also an argument that more consideration should be devoted to the conduct of future systematic reviews. Building upon these issues it is suggested that as part of the systematic review process, consideration should be given to omitting studies that are not regarded as scientifically admissible, the definition of which is taken from current literature and expert opinion where needed. Pre-definition of what constitutes scientific admissibility in a research study offers a structured and transparent approach in contrast to other methods currently in use. For example, in relation to the RCT current thinking suggests that type of allocation, blinding and appropriate sample size are key design features that when absent or compromised might introduce bias into the results of RCT's (Kjaergard et al

2001). Other features, including the validity of outcome measures, may warrant consideration before inferring that studies are scientifically admissible as part of the review process. Hence rather than simply considering the types of studies, participants, interventions, comparisons and outcomes (PICO) as a means of including studies we are suggesting that the key components of the methodology that would contribute to scientific admissibility should form part of the inclusion and exclusion criteria of the review, hence the acronym PICOM. This addition recognises the impact that research methods might have and it is suggested that the addition of this process would add credibility to the findings of systematic reviews and hence may be a stronger basis upon which to develop clinical practice. As with conventional reviews, it is at this point where studies have been identified as suitable for inclusion that further quality appraisal could be undertaken, using established scales, to indicate the relative quality of the included studies.

This modification to the process would seem to complement contemporary systematic review methods for which there is a plethora of guidance available regarding how to conduct a high quality systematic review, including the Cochrane Handbook (Higgins and Green 2009) and the NHMRC guidelines (1999), and hence it seems unnecessary in a paper of this nature to repeat this work.

## **Conclusion**

In light of the limitations associated with current systematic review methods and the different conclusions that these methods may deliver it seems that there is an argument emerging for the physiotherapy profession to consider re-evaluating the status of its research base by conducting further systematic reviews which omit low quality studies with a high potential for bias, reduce reliance on numerical rating scales to infer quality whilst offering a current, robust and transparent approach to quality appraisal and data synthesis as part of the systematic review process.

## **References**

Assendelft W, Koes B, Knipschild P, Bouter L, (1995). The relationship between methodological quality and conclusions in reviews of spinal manipulation. *Journal of the American medical association*, 274; 1942-1948.

Barr S, Cerisola F, Blanchard V, (2009). Effectiveness of corticosteroid injections compared with physiotherapy interventions for lateral epicondylitis: a systematic review. *Physiotherapy*, 95; 251-265.

Carroll L, et al (2008). Methods for the best evidence synthesis on neck pain and its associated disorders. The Bone & Joint Decade 2000-2010 Task Force on Neck Pain and Its Associated Disorders. *Spine*, 30; S33-S38.

Hettinga D, Hurley D, Jackson A, May S, Mercer C, Roberts L, (2008). Assessing the effect of sample size, methodological quality and statistical rigour on outcomes of randomised controlled trials on mobilisation, manipulation and massage for low back pain of at least 6 weeks duration. *Physiotherapy*, 94; 97-104.

- Higgins J, Green S, (Eds), (2009). The Cochrane Handbook for systematic reviews of interventions: Version 5.0.2. Last accessed 27th May 2010 via: <http://www.cochrane-handbook.org>.
- Ho C, Sole G, Munn J (2009). The effectiveness of manual therapy in the management of musculoskeletal disorders of the shoulder. A systematic review. *Manual therapy*, 14; 463-474.
- Kjaergard L, Villumsen J, Gluud C (2001). Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. *Annals of internal medicine*, 135; 982-989.
- LeLorier J, Gregoire G, Benhaddad A, Lapierre J, Derderian F (1997). Discrepancies between meta-analyses and subsequent large randomized controlled trials. *The New England journal of medicine*, 337; 536-542
- Littlewood C, May S (2007) Measurement of range of movement in the lumbar spine - what methods are valid? A systematic review. *Physiotherapy*, 93; 201-211.
- Maher C, Sherrington C, Herbert R, Moseley A, Elkins M (2003). Reliability of the PEDro scale for rating quality of randomized controlled trials. *Physical therapy*, 83; 713-721.
- May S, Littlewood C, Bishop A (2006). Reliability of procedures used in the physical examination of non-specific low back pain: A systematic review. *Australian journal of physiotherapy*, 52; 91-102.
- Moore R, Gavaghan D, Tramer M, Collins S, McQuay H (1998). Size is everything - large amounts of information are needed to overcome random effects in estimating direction and magnitude of treatment effects. *Pain*, 78; 209-216.
- National Health and Medical Research Council (NHMRC), (1999). How to review the evidence: systematic identification and review of the scientific literature. Australia: NHMRC.
- Paratz J, Stockton K (2009) Efficacy and safety of normal saline irrigation: a systematic review. *Physiotherapy*, 95; 241-250.
- Slavin R (1995). Best evidence synthesis: An intelligent alternative to meta-analysis. *Journal of clinical epidemiology*, 48; 9-18.
- Tang C, Taylor N, Blackstock F (2010). Chest Physiotherapy for patients admitted to hospital with an acute exacerbation of chronic obstructive pulmonary disease (COPD): a systematic review. *Physiotherapy*, 96; 1-13.
- Van der Velde G, van Tulder M, Cote P, Hogg-Johnson S, Aker P, Cassidy D (2007). The sensitivity of review results to methods used to appraise and incorporate trial quality into data synthesis. *Spine*, 32; 796-806.
- Van Trijffel E, Anderegg Q, Bosuyt P, Lucas C (2005). Inter-examiner reliability of passive assessment of intervertebral motion in the cervical and lumbar spine: A systematic review. *Manual therapy*, 10; 256-269.

Van Tulder M, Furlan A, Bombardier C, Bouter L (2003). Updated method guidelines for systematic reviews in the Cochrane Collaboration Back Review Group. *Spine*, 28; 1290-1299.