ORIGINAL ARTICLES

PEDro’s bias: summary quality scores should not be used in meta-analysis

Bruno R. da Costa, Roger Hilfiker, Matthias Egger

Institute of Social and Preventive Medicine, University of Bern, Finkenhubelweg 11, 3012 Bern, Switzerland
Department of Physiotherapy, University of Applied Sciences Western Switzerland, Sion, Switzerland

Accepted 20 August 2012

Systematic reviews of randomized controlled trials and statistical combination of results from different trials in meta-analysis have risen to such prominence that few question their usefulness in evidence-based clinical practice. Although systematic reviews are indeed generally useful, the results of meta-analyses must be cautiously interpreted. If the estimates of the trials included in the meta-analysis deviate from the truth in a systematic fashion, the summary estimate from the meta-analysis will likely be biased as well.

There is strong empirical evidence showing that the meta-analyses of clinical trials in which the allocation of patients to treatment groups was not concealed, or in which the assessment of outcomes was not blinded, overestimate the treatment effects. It is now widely accepted that the quality of a trial should be assessed before including it in a meta-analysis. Likewise, it is a good practice to ascertain if the results differ between trials at greater or lesser risk of bias. As yet, however, there is no consensus as to how this assessment should be done. In this commentary, we discuss the scoring of trials using quality scales, using the example of a scale widely used in trials of physiotherapy. We contrast the use of summary scores with an alternative approach, which is based on an assessment of individual components such as concealment of allocation and blinding.

Quality scales assess several criteria related to the design, conduct, and analysis of trials, and each earns points that are aggregated into an overall score. The score determines the classification of the study as one of the higher or lower methodological quality, with the implication that bias has been prevented to a greater or lesser degree. Chalmers et al. in 1981 were among the first to develop such a scale, with possible summary scores ranging from 0 to 44. Since then, many other scales were developed, including the widely used Physiotherapy Evidence Database (PEDro) scale [7–10]. Table 1 details the 11 items included in the PEDro scale. Most relate to the design and conduct of the trial, but three are concerned with reporting eligibility criteria (item 1), between-group statistical comparisons (item 10), and measures of variability (item 11). Notably, only two of the three items on reporting quality contribute points to the total score: “eligibility criteria specified” does not. The total score is therefore 10 rather than 11. Although the PEDro scale was developed for clinical trials of physiotherapy, it does not contain items that are specific to this field. For example, efforts to reduce bias owing to differential expertise of therapists are not considered by the PEDro scale. A physiotherapist’s training, experience, and personal preferences are likely to influence the efficacy of randomized therapies. Such bias may be reduced if patients are randomly allocated to practitioners who have expertise in the intervention under investigation.

The PEDro scale is routinely used to assess the quality of the randomized controlled trials that are continuously added to the freely available Physiotherapy Evidence Database [7,8] (also see http://www.pedro.org.au/). The database currently covers more than 15,000 clinical trials and makes their quality scores available. The PEDro scale is also widely used in published systematic reviews and meta-analyses. A search of Medline in March 2012 identified 126 reports that mentioned the PEDro scale in the abstract. Twenty-three used the PEDro score as an eligibility criterion, or to stratify analyses by trial quality, with the threshold score for eligibility or high quality ranging from three to seven. Of note, the database allows searches to be restricted to trials with a minimum PEDro score defined by the user and thus encourages the exclusion of trials below a given threshold.

The Cochrane Collaboration recommends a component approach to the assessment of trial quality and risk of bias, rather than the use of quality scales such as PEDro [12]. It was wondered whether the approach chosen might affect the conclusions of Cochrane reviews. To examine this, the
review of transcutaneous electrostimulation for osteoarthritis of the knee was revisited [13]. The PEDro scores were available for 15 of the 16 trials included in the meta-analysis, with scores ranging from three to eight. These 15 trials contributed 17 comparisons. The Cochrane review had identified only one trial of clearly high quality [14], with adequate generation of random sequence, concealment of allocation, and blinding of study participants and therapists. This trial received a high PEDro score of eight; however, other trials that were also highly scored by PEDro lacked adequate concealment of allocation and blinding of patients or therapists [15–19]. A high PEDro score does thus not mean that the trial was adequately randomized and blinded.

The Cochrane meta-analysis was repeated [13] using the scores from the PEDro database to identify high-quality trials. Fig. 1 shows the summary effect sizes for pain for all trials, for trials of higher quality using different PEDro thresholds, and for the high-quality trial identified in the original review [13]. With the PEDro scale, the beneficial effect of electrostimulation became more prominent as the quality of the trials increased. In contrast, the estimate from the high-quality trial was close to zero, indicating little benefit of electrostimulation. The previous review concluded that the evidence was “inconclusive, hampered by the inclusion of only small trials of questionable quality” [13]. It seems likely that based on the results from the PEDro scale, many reviewers would conclude with the contrary opinion that there was robust evidence from high-quality trials that electrostimulation had a clinically relevant, beneficial effect on pain in osteoarthritis of the knee.

Greenland [20] described quality scoring as “perhaps the most insidious form of subjectivity masquerading as objectivity in meta-analysis”: the effects of quality dimensions that are important in a given study and context are diluted or confounded by the summary quality score, sometimes to the point that quality effects are no longer evident, or that effects are reversed, as in our example. The PEDro scale and many other quality scales include items that are not in fact related to the methodology and extent bias was avoided in a trial, but to the quality of reporting. Furthermore, items that are important for some interventions or outcomes may not be relevant in other situations but will receive the same consideration. For example, blinding of study participants will be crucial for pain but irrelevant for all-cause mortality. Greenland’s arguments are supported by our case study and other empirical evidence. Jüni et al. [22] used 25 different scales to assess the quality of trials comparing low—molecular weight heparin with standard heparin. In 25 separate meta-analyses (one for each quality scale), the results from high- and low-quality trials were compared, with very different results depending on which quality scale was used. Colle et al. [21] found similar effects across 16 different scales for trials of exercise for low back pain. Finally, Wood et al. [4] showed in a study of 146 meta-analyses that blinding was associated with exaggerated effects in trials with subjective outcomes, but not in trials with objective outcomes.

The assembly of the PEDro database of many thousands of trials, dating back to the 1930s, is a tremendous achievement and an extremely valuable resource for the evaluation of the evidence underpinning physiotherapy interventions. Unfortunately, the PEDro database’s inappropriate emphasis on the use of summary scores from a quality scale makes the database less useful than it might be. It is likely that bias is introduced into systematic reviews and meta-analyses.

Table 1. Items of the Physiotherapy Evidence Database scale

<table>
<thead>
<tr>
<th>Item No.</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Eligibility criteria were specified</td>
</tr>
<tr>
<td>2</td>
<td>Subjects were randomly allocated to groups</td>
</tr>
<tr>
<td>3</td>
<td>Allocation was concealed</td>
</tr>
<tr>
<td>4</td>
<td>The groups were similar at baseline regarding the most important prognostic indicators</td>
</tr>
<tr>
<td>5</td>
<td>There was blinding of all subjects</td>
</tr>
<tr>
<td>6</td>
<td>There was blinding of all therapists who administered the therapy</td>
</tr>
<tr>
<td>7</td>
<td>There was blinding of all assessors who measured at least one key outcome</td>
</tr>
<tr>
<td>8</td>
<td>Measures of at least one key outcome were obtained from more than 85% of the subjects initially allocated to groups</td>
</tr>
<tr>
<td>9</td>
<td>All subjects for whom outcome measures were available received the treatment or control condition as allocated or, where this was not the case, data for at least one key outcome was analyzed by “intention to treat”</td>
</tr>
<tr>
<td>10</td>
<td>The results of between-group statistical comparisons are reported for at least one key outcome</td>
</tr>
<tr>
<td>11</td>
<td>The study provides both point measures and measures of variability for at least one key outcome</td>
</tr>
</tbody>
</table>

Item 1 (eligibility criteria) does not contribute to total score.

Fig. 1. Standardized mean differences (SMDs) in pain scales from trials of transcutaneous electrostimulation for osteoarthritis of the knee. Trials of high-methodological quality are either defined based on PEDro summary scores or adequate randomization and blinding. The SMDs were combined in inverse-variance random-effects meta-analysis. Adapted from a Cochrane review [13]. PEDro, Physiotherapy Evidence Database.
when these scores are used as the main criteria on which the inclusion or exclusion of trials is based. We suggest that the use of summary scores should be discouraged, and that the PEDro database be restricted to presenting the scores for individual items of the scale. The addition of items, for example on the prevention of differential expertise bias, and the removal of items not related to the risk of bias, might further enhance the value of this important initiative.

References


