ELSEVIER

Journal of Clinical Epidemiology 61 (2008) 209–231

Journal of Clinical Epidemiology

REVIEW ARTICLES

Methodological quality and homogeneity influenced agreement between randomized trials and nonrandomized studies of the same intervention for back pain

Andrea D. Furlan[a,b,*], George Tomlinson[b,c], Alejandro (Alex) R. Jadad[b,d,e], Claire Bombardier[a,b,c,f,g]

[a]*Institute for Work & Health, Toronto, ON, Canada M5G 2E9*
[b]*University of Toronto, Department of Health Policy, Management and Evaluation, Canada*
[c]*Department of Medicine at University of Toronto, Toronto General Research Institute, Canada*
[d]*Centre for Global eHealth Innovation, Canada*
[e]*Rose Family Chair in Supportive Care, Canada*
[f]*Division of Rheumatology, University of Toronto, Canada*
[g]*Mount Sinai Hospital, Toronto, Canada*

Accepted 21 April 2007

Abstract

Objective: To determine the influence of methodological quality and homogeneity on the agreement between pairs of randomized trials (RCTs) and nonrandomized studies (NRSs) of the same interventions for low-back problems. Homogeneity was assessed regarding settings, population, interventions, and outcomes.

Study Design and Setting: We searched Cochrane Central Register of Controlled Trials, MEDLINE, and EMBASE up to May 2005 for matching pairs of NRS and RCT. Analyses were done using correlation, linear and logistic regression.

Results: Forty-eight matched pairs were included with no significant overall correlation between effect sizes ($r = 0.09$). There was a trend showing more agreement among the 22 pairs with higher methodological quality ($r = 0.33$). The correlation among the 20 very homogeneous pairs was 0.59, and among the 28 heterogeneous pairs was −0.09. The agreement of authors' recommendations was influenced by the pair's homogeneity (odds ratio [OR] = 2.78, 95% CI = 1.44–5.37) rather than by methodological quality of the NRS (OR = 0.93, 95% CI = 0.67–1.29) or the RCT (OR = 1.03, 95% CI = 0.73–1.45).

Conclusions: Pairs of low-quality studies disagreed more than pairs where at least one study was of high quality. However, pairs with similar settings, population, interventions, and outcomes showed higher agreement than pairs that were not as homogeneous.

Keywords: Systematic review; Randomized controlled trial; Observational study; Cohort study; Effect sizes; Low-back pain

1. Introduction

The number of primary studies assessing the effects of health care interventions is continuously growing. This might create a problem when authors from different studies reach conflicting conclusions about the effectiveness of a determined health care intervention. Systematic reviews are supposed to solve these conflicts by employing explicit methods to search, select, critically appraise, and combine all available primary studies on a prespecified topic, thereby making objective recommendations about the effects of these interventions. Among all kinds of primary studies, randomized controlled trials (RCTs) are considered the most reliable method to determine the efficacy and effectiveness of health care interventions because they can minimize biases [1]; however, conflicting conclusions among RCTs are not uncommon. Furukawa et al. examined 289 pairs of megatrials (>1,000 subjects) that studied the same patient–intervention–outcome combinations and found that in 53% of pairs the treatment recommendations in the first trial were not confirmed by a later megatrial. They explain that "this much agreement or disagreement among megatrials may in fact be reasonable, given the

* Corresponding author. Tel.: 1-416-927-2027; fax: 1-416-927-4167.
E-mail address: afurlan@iwh.on.ca (A.D. Furlan).

0895-4356/08/$ – see front matter
doi: 10.1016/j.jclinepi.2007.04.019

What is new?

1. Agreement between pairs of randomized trials and nonrandomized studies (26 out of 48 pairs) was most common when the pairs were of higher methodological quality.
2. Comparison between a randomized trial and a nonrandomized study should take into consideration the methodological quality of the studies in addition to the clinical homogeneity of the pair, in terms of settings, population, interventions, and outcomes.
3. This "homogeneity index" should be used for future comparisons or interpretations of past comparisons between randomized trials and nonrandomized studies.

inevitable variations in the patients involved, in the way interventions are administered, in the measurement of the outcomes, and the design of studies among megatrials" [2].

Many systematic reviews do not include evidence from nonrandomized studies (NRSs) when assessing the effectiveness of health care interventions because of the assumption that their results always overestimate the effects of that intervention [3]. A number of recent studies have compared the results of RCTs and NRSs. In 2000, Concato et al. showed that NRSs do not systematically overestimate the magnitude of the effects of treatment as compared with those in RCTs on the same topic [4]. Also in 2000, Benson and Hartz found little evidence that estimates of treatment effects in observational studies reported after 1984 were either consistently larger than or qualitatively different from those obtained in RCTs [5]. In the field of low-back pain, Furlan et al. showed that NRSs either agreed with or underestimated the effects of various surgical interventions compared to RCTs [6].

In 1998, Britton et al. suggested that RCTs and NRSs can produce different magnitude and direction of results, but variation in results also occurs among RCTs and among NRSs. They concluded that these differences often reflect variations in design features that may be sufficiently different to preclude detection of any specific effect of the process of allocation [7]. There has been no study showing that conflicts between RCTs and NRSs could be explained on the basis of different settings, population, interventions, or outcomes.

Our hypotheses were that 1) agreement is higher among pairs of higher methodological quality studies compared to pairs of lower quality studies, and 2) that RCTs and NRSs that share many clinical similarities would agree more than RCTs and NRSs that have many clinical differences. This study may help guide the design of future comparisons of RCTs and NRSs or the interpretation of past comparisons. It may be useful for developing guidelines for the minimal similarity between an RCT and an NRS before a comparison of the effect sizes is meaningful.

2. Methods

2.1. Search and selection of nonrandomized studies

Comprehensive searches were conducted in MEDLINE and EMBASE up to June 2002 and updated in May 2005. The complete search strategies used can be obtained from the corresponding author. One person (A.D.F.) ran the searches and selected the studies based on the following inclusion criteria:

- Nonrandomized study comparing two or more interventions for low-back pain. The taxonomy of study design shown in Table 1 was used. Only observational studies with a comparison group were included;
- Any intervention (prevention or treatment);
- Any type of low-back pain (acute or chronic, nonspecific etiology, disc herniation, spinal stenosis, osteoarthritis, etc.);
- Any type of outcomes measure.

2.2. Matching phase

The nonrandomized studies were grouped according to the type of intervention. One author (A.D.F.) searched for RCTs for each intervention by: 1) searching reference lists of the Cochrane Reviews and non-Cochrane systematic reviews of the same intervention, 2) searching the Cochrane Central Register of Controlled Trials using keywords related to the intervention, and 3) searching MEDLINE and EMBASE using a simple search strategy that is shown to be highly sensitive [8].

One author (A.D.F.) performed the matching of NRSs and RCTs when the two studies included similar populations, interventions, comparison groups, and outcome measures.

2.3. Quality ratings of nonrandomized studies

Two independent reviewers (A.D.F. and Mana Rezai) rated the methodological quality of all 48 NRSs, using consensus to solve disagreements. Agreement between the two reviewers was measured using kappa statistics for each item (yes or no) and for the summary score (low or high quality). Intraclass correlation coefficient was used to measure agreement for the total score (range: 0–12).

We used the 12 questions related to the six domains suggested by Deeks et al. to rate internal validity of nonrandomized intervention studies (Table 2) [9]. Deeks et al. identified 194 tools that could be or had been used to assess NRSs. They used a modified Delphi process