The art of quality assessment of RCTs included in systematic reviews

Arianne P. Verhagen a,b,*, Henrica C.W. de Vet c,d, Robert A. de Bie a,
Maarten Boers e, Piet A van den Brandt e

 a, b Department of Epidemiology, Maastricht University, Maastricht, The Netherlands
 c Department of General Practice, Erasmus University Rotterdam, Rotterdam, The Netherlands
 d Institute of Extramural Medicine, Vrije Universiteit, Amsterdam, The Netherlands
 e Department of Clinical Epidemiology, Vrije Universiteit, University Hospital, Amsterdam, The Netherlands

Received 4 January 2000; received in revised form 3 October 2000; accepted 23 October 2000

Abstract

The best evidence on the efficacy of medical interventions is provided by high-quality trials summarized in high-quality systematic reviews or meta-analyses. The methodological quality of studies included in a systematic review can have a substantial impact on the estimates of the treatment effect and therefore on the conclusions of such a review. But what is the empirical evidence to support quality assessment of randomized clinical trials (RCTs)? We elaborate on questions such as: what is the concept of quality of individual studies (RCTs), can quality be measured validly and reliably? Plans for future research on this issue are proposed. © 2001 Elsevier Science Inc. All rights reserved.

1. Introduction

Well-conducted randomized clinical trials (RCTs) provide the best evidence on the efficacy of medical interventions. In 1998 the British Medical Journal celebrated the 50th anniversary of the RCT in the medical field. During the past 50 years the number of RCTs published yearly has increased immensely. According to MEDLINE in 1997 more than 29,000 new RCTs were published. For the ordinary clinician it has become impossible to keep up with the latest evidence.

Systematic reviews are designed to help the clinician base their clinical decisions on the best available evidence. These reviews include a comprehensive search strategy and a predetermined and explicit method to appraise and synthesize the information from individual studies [1,2]. Systematic reviews are a form of observational research and therefore susceptible to bias. In the conduct of a systematic review or meta-analysis, many possible sources of bias exist, such as publication bias (are all published and unpublished studies included); heterogeneity (differences in study population, interventions, outcome measures, etc.); language bias (were we able to include reports written in non-English languages); and bias caused by design characteristics. All are considered having influence on the reported outcome of a systematic review. Apart from these factors the validity of the conclusions of a systematic review will obviously depend on the quality of the included primary studies. Therefore, assessment of trial quality is often a part of the process of a systematic review. But what is the empirical evidence to support quality assessment of RCTs? In this article we focus on the empirical evidence of quality assessment as one of the elements in the protocol of a systematic review.

2. What is quality?

Most criteria lists proposed to assess the methodological quality of RCTs do not explicitly define the concept of quality [3]. These lists usually include at least three dimensions that may encompass the concept of quality: internal validity, external validity and statistical analysis [4–6]. Quality of RCTs has recently been defined as: "the likelihood of the trial design to generate unbiased results" [7]. This definition covers only the dimension of internal validity. During the development of the "Delphi list" for quality assessment, the participants, all experts in the field of RCTs, failed to reach consensus on a specific definition, but did agree that the concept of quality should comprise more than internal validity alone [8]. From this context we propose the following definition of quality: the likelihood of the trial design to generate unbiased results, that are sufficiently precise and allow application in clinical practice.

* Corresponding author. Department of General Practice, Erasmus University Rotterdam, P.O. Box 1798, 3000 DR Rotterdam, the Netherlands. Tel.: +31 10 408 8109; fax: +31 10 408 9491.
E-mail address: Verhagen@hag.fgg.eur.nl (A.P. Verhagen)

0895-4356/01 - see front matter © 2001 Elsevier Science Inc. All rights reserved.
Pll: S0895-4356(00)00360-7
3. How can quality be assessed?

The last decade has shown efforts to develop appropriate tools for quality assessment. This includes the 1996 CONSORT guidelines [9] that aim to set the standard for the written report of an RCT. When assessing trial quality one has to rely on the information retrieved from these reports. One approach in quality assessment is to focus on components such as randomization, blinding etc. in trial reports [10,11]. Another is to use a criteria list to provide a quality score as an estimation of the overall methodological quality of the design and conduct of the trial. Higher quality scores should indicate studies with a better methodological quality [12,13]. Moher et al. reported in 1995 that at least 24 criteria lists have been developed [3]. We currently estimate the number of quality scales between 50 and 60, and the number is still increasing.

4. What is the evidence for the validity and reliability of quality assessment?

The validity of quality assessment itself needs further study. Such studies should answer the question: does this criteria list measure what it is supposed to measure, namely the methodological quality of the trial? To know how close a criteria list measures the true state, we need a gold standard or external criterion to compare it against (criterion validity). However, a gold standard of quality assessment does not, and probably will never, exist. In such cases one has to fall back on a theoretical model. This means assessing the content validity: does the method of measurement include all dimensions of the theoretical framework? Face validity: does the criteria list seems valid when looked at by experts? Most tools to assess trial quality have a reasonable face validity or content validity because they comprise a selection of "accepted criteria," such as those listed in textbooks on clinical trials as aspects of importance for the quality of a trial [3].

We feel that the application of formal scale development techniques, including consensus, can increase the validity of the resulting quality scale. As far as we are aware the only criteria lists developed by such techniques are the list developed by Jadad et al. [7] and the Delphi list [8]. Although consensus provides some validity, it is not a paradigm. Consensus is always achieved within a theoretical model, which can be proven wrong in time. Notably, consensus did not prevent marked differences between the Jadad and the Delphi list. Whether quality assessment is of real value in the process of a systematic review partly depends on the validity and reliability of the criteria list used.

Another aspect of validity is construct validity: is the measurement consistent with other measurements of quality assessment? This means comparing the results of different quality criteria lists with each other. Recently, Jüni and colleagues [14] compared the results of 25 criteria lists on trials comparing low-molecular weight heparin with standard heparin for thromboprophylaxis in general surgery. Their conclusion was that the conclusion of a meta-analysis, when based on only high-quality trials, differed depending on which criteria list was used. Does this result mean that none of the quality criteria lists are valid? The validity of most of the 25 criteria lists used is related to the topic of the research. For instance, can criteria lists designed for acupuncture [15] or homeopathy trials [16] validly measure the quality of trials on low-molecular weight heparin? This study of Jüni et al. [14] is one, very important, empirical case study and many more similar studies need to be performed in order to evaluate the construct validity of quality assessment.

The reliability of most criteria lists is unknown. In reviews where more than one reviewer assesses the trial quality, reviewers often discuss their differences and reach a consensus score. When studied, the inter-rater agreement assessed for several criteria lists vary from moderate to good [7,17–20]. Intra class correlation coefficients (ICC) were calculated by Jadad et al. [7] and Bender et al. [18] for the Jadad list and by Verhagen et al. [19] for the Maastricht list. These ICCs varied from 0.65 to 0.85, all regarded as "high." Recently, Clark et al. [20] calculated an inter-rater agreement on the Jadad list using Kappa values. These appeared to be low (0.37–0.39). Emerson et al. [17] calculated for his criteria list an overall agreement, which appeared to be roughly 80%.

5. What are major potential biases in the application of quality assessment?

Reporting bias is a specific form of information bias: are published RCTs a true reflection of what went on in the trial? Reporting of trials may be flawed in a way that provides a misleading impression of methodological quality, one way or the other [21]. On one hand, poorly reported trials could be judged as having low quality, while it may not be so. This is a problem of "under-reporting." A flawed report (i.e., lacking the necessary information about trial design and conduct) does not necessarily mean that the underlying study was flawed. It may reflect a lack of understanding of the reporting requirements for such studies [22]. On the other hand, quality can also be "over-reported." Between 5% and 30% of "randomized" studies may actually not have performed a method of randomization [23,24]. With the growing emphasis on the quality of trials, the problem of over-reporting may be increasing. A recent initiative of the BMJ and Lancet [25,26] to promote a register of trials might partly overcome the problem of reporting bias. The main aim of this register is to reduce publication bias, but comparing the original trial report with the actual manuscript may also provide insight in possible over- or under-reporting.

Review bias is a type of information bias specifically for reviews. Beliefs and disbelief's can (subconsciously) guide
reviewers into biased assessments. When reviewers know, for example, that a study is published in a high-impact journal, they might unconsciously rate the study more positive (i.e., of higher quality), because they trust the peer review process of that journal. Another possible cause of review bias could occur when the profession of the reviewers is linked to the intervention investigated (clinicians) rather than epidemiologists as reviewers [19]. To prevent review bias, it has been suggested to perform quality assessment under masked conditions (i.e., authors, institutes, sponsorships, journals of publication, or study results should be unknown to the reviewer) [5, 7, 27–29]. To date however, research concerning the necessity of masked quality assessment has shown no consistent results [7, 19, 20, 30].

Bias due to misclassification may result when overall quality is used to determine a cut-off point between high-versus low-quality studies. In the case of exclusion of valid studies from the systematic review, the precision is most likely to be reduced, but inclusion of invalid studies leads almost certainly to biased results. When the validity and reliability of the quality criteria list is unknown, there is a real chance of bias due to misclassification [14]. Using only quality components might decrease the problem of misclassification, but does not fully overcome it.

6. How can we incorporate quality into the conclusion?

Whether or how the results of quality assessment should be incorporated into the conclusion of a review is under debate, especially when quality scores are used [31, 32]. Several strategies are available to do this [32]. First, a visual plot of the effect size against an overall quality score can be presented [10, 32–34]. Further, quality components or scores can be used as a “threshold score” for inclusion of the article in a review, as a “weighting factor” in the statistical analysis, [10, 32, 35] or as the input sequence in a cumulative meta-analysis [10, 33, 34]. Finally, meta-regression techniques are proposed to study the impact of various quality components on the conclusion of the meta-analysis [31, 36]. This latter method is only possible in reviews containing a large number of trials. In other situations a visual plot of effect size against quality components or scores provides most insight into whether and how quality influences the final conclusion.

7. Which way ahead: quality components or quality scores?

The leading paradigm in the field of quality assessment of RCTs is that low-quality studies tend to overestimate effect estimates. This paradigm is based on a theoretical framework, which assumes that investigators are (subconsciously) biased in favor of the intervention [11]. Empirical research has shown that components of quality can indeed influence the effect estimates; however, the direction of this influence is not consistent [11, 21, 30, 37]. In other words low study quality can both underestimate and overestimate the true effect [37].

In our opinion empirical research should primarily focus on components of quality, measured using a criteria list. Apart from randomization and blinding items, this criteria list should contain items concerning other design characteristics that possibly influence the results (e.g., the Delphi list) [8]. Empirical research should determine the relevance of these and other items. Such research should then guide the improvement of already existing quality criteria lists.

8. Plans for the future

We believe quality assessment of randomized clinical trials is essential in conducting meaningful systematic reviews, but the quality of this exercise is yet unclear. One of the advantages of systematic reviews is that all the choices reviewers make in the review process concerning eligibility of studies, search strategies, quality assessment, etc., are transparent when explicitly described in the manuscript. This enables readers to judge for themselves the adequacy of quality assessment and potential sources of bias concerning quality assessment in the conduct of the review.

The Cochrane Collaboration publishes guidelines on how to perform systematic reviews, but methodological studies to evaluate the relationship between quality or design characteristics and effect sizes are needed. Examples of such empirical studies are studies concerning the construct validity (such as the study of Jüni et al. [14]), studies concerning the reliability of several quality assessment instruments (such as the study of Clark et al. [20]), and studies concerning the impact of design characteristics on outcome (such as Linde et al. [36]). In our opinion future studies should primarily focus on design characteristics. These design characteristics should preferably be measured in a uniform way in several different reviews. Emphasis should be put on two or three generic criteria lists for quality assessment, and to study the impact of quality components. We strongly recommend performing future research within a specific research question in order to minimize heterogeneity of, for instance, the study population and outcome measures, as we assume that the relevance of several items will strongly depend on the topic of research. These studies can eventually qualitatively be summarized into a methodological review. This way it might be possible to determine empirically which items in these generic criteria lists are important and have an impact on outcome. Cochrane review groups can always add specific items to one of these generic criteria lists, relevant for the aim of the review.

9. Epilogue

We consider quality assessment a valuable theoretical model. In this model differentiation in quality is essential in order to find a valid and clinically relevant effect estimate in systematic reviews. The task for future research is to gener-
ate a valid set of quality criteria to provide more insight in the still hazy relationship between quality and outcome. In the meantime, a detailed description of the quality assessment procedure and other choices made in the conduct of the review should enable clinicians and other readers to judge the validity of the procedure. This description should include the quality components that are assessed, the way these are used in the analysis (as components or scores), and the way they may influence the results. Plots of the relationship between effect size and quality measures can visualize the latter, by meta-regression or sensitivity analysis. Clinicians can benefit from information concerning quality assessment in a review in a way that he/she is now able to determine whether there is sufficient (high-quality) evidence to prescribe a certain therapy or medication to a specific patient.

References


