Systematic reviews in health care
Assessing the quality of controlled clinical trials
Peter Juni, Douglas G Altman, Matthias Egger

The quality of controlled trials is of obvious relevance to systematic reviews. If the “raw material” is flawed then the conclusions of systematic reviews cannot be trusted. Many reviewers formally assess the quality of primary trials by following the recommendations of the Cochrane Collaboration and other experts. However, the methodology for both the assessment of quality and its incorporation into systematic reviews and meta-analysis are a matter of ongoing debate. In this article we discuss the concept of study quality and the methods used to assess quality.

Quality is a multidimensional concept, which could relate to the design, conduct, and analysis of a trial, its clinical relevance, or quality of reporting. The validity of the findings generated by a study clearly is an important dimension of quality. In the 1950s the social scientist Campbell proposed a useful distinction between internal and external validity (see box below). Internal validity is threatened by bias—any process at detection bias, performance bias, and attrition bias among others—but is minimised in clinical trials. Selection bias, biased allocation to comparison groups, and performance bias, unequal provision of care apart from treatment under evaluation, are amongst the problems that can occur. Attrition bias, biased occurrence and handling of deviations from protocol and loss to follow up cause external validity to be threatened. Internal validity becomes redundant when the results of a study provide a correct basis for generalisation to other circumstances. In itself, there is no external validity. The term is only meaningful with regard to specified “external” conditions, such as other patient populations or treatment regimens. Internal validity is a prerequisite for external validity: the results of a flawed trial are invalid, and the question of its external validity becomes redundant.

Components of internal and external validity of controlled clinical trials

**Internal validity**—extent to which systematic error (bias) is minimised in clinical trials

- Selection bias: biased allocation to comparison groups
- Performance bias: unequal provision of care apart from treatment under evaluation
- Detection bias: biased assessment of outcome
- Attrition bias: biased occurrence and handling of deviations from protocol and loss to follow up

**External validity**—extent to which results of trials provide a correct basis for generalisation to other circumstances

- Patients: age, sex, severity of disease and risk factors, comorbidity
- Treatment regimens: dosage, timing and route of administration, type of treatment within a class of treatments, concomitant treatments
- Settings: level of care (primary to tertiary) and experience and specialisation of care provider
- Modalities of outcomes: type or definition of outcomes and duration of follow up

Summary points

Empirical studies show that inadequate quality of trials may distort the results from systematic reviews and meta-analyses

The influence of the quality of included studies should routinely be examined in systematic reviews and meta-analyses

The use of summary scores from quality scales is problematic—it is preferable to examine the influence of key components of methodological quality individually

Based on empirical evidence and theoretical considerations, the generation and concealment of the allocation sequence, blinding, and handling of patient attrition in the analysis should always be assessed

Dimensions of internal validity

Internal validity is threatened by bias, “any process at any stage of inference tending to produce results that differ systematically from the true values.” In clinical trials, biases fall into four categories: selection bias, performance bias, detection bias, and attrition bias (box).
Selection bias
The aim of randomisation is the creation of groups that are comparable for any known or unknown potential confounding factors. Success depends on two interrelated procedures (see above). Firstly, an allocation sequence that is suitable to prevent selection bias must be generated—for example, by using a computer algorithm, tossing a coin, or throwing a dice. Secondly, this sequence must be concealed from investigators enrolling patients. Knowledge of assignments—for example, from a table of random numbers posted on a bulletin board—can cause selective enrolment of patients on the basis of prognostic factors. Patients who have been assigned to a treatment deemed to be “inappropriate” may be rejected, and some patients may be deliberately directed to the “appropriate” treatment. Deciphering of allocation schedules may occur even if concealment was attempted. For example, envelopes may be opened or held against a bright light to reveal the contents.

Performance bias and detection bias
Performance bias occurs if additional treatment interventions are provided preferentially to one group. Blinding of patients and care providers prevents this type of bias and also safeguards against differences in placebo responses between the groups. Detection bias arises if the knowledge of patient assignment influences the assessment of outcome. This is avoided by the blinding of those assessing outcomes—for example, patients, care providers, radiologists, or end point review committees (box).

Attrition bias
Deviations from protocol and loss to follow up often lead to the exclusion of patients after they have been allocated to treatment groups, which may introduce attrition bias. Possible deviations from protocol include the violation of eligibility criteria and non-adherence to treatments. Loss to follow up refers to patients becoming unavailable for examinations at some stage during the study period because they refuse to participate further (also called drop outs), cannot be contacted, or clinical decisions are made to stop the assigned interventions.

Patients excluded after allocation are unlikely to be representative of patients remaining in the study. For example, patients may not be available for follow up because they have an acute exacerbation of their illness or severe side effects. Patients not adhering to treatments generally differ in respects that are related to prognosis. All randomised patients should therefore be included in the analysis and kept in their original groups, regardless of their adherence to the study protocol. In other words the analysis should be performed according to the intention to treat principle, thus avoiding selection bias. This implies that the primary outcome was recorded for all randomised patients at the prespecified times throughout the follow up period. If the end point of interest is mortality from all causes this can be established most of the time. It may, however, be impossible retrospectively to ascertain other binary or continuous outcomes, and some patients may therefore have to be excluded from the analysis. In this case the proportion of patients not included in the analysis must be reported and the possibility of attrition bias discussed.

Empirical evidence of bias
Numerous case studies show that the biases described above do occur in practice, distorting the results of clinical trials. The authors are aware of four methodological studies that have gauged their relative importance in a large number of clinical trials while avoiding confounding by disease or intervention. The figure shows a meta-analysis of the results from these studies. Inadequate or unclear concealment of treatment allocation was associated with an exaggeration of treatment effects in all four studies. Odds ratios from trials with inadequate or unclear concealment were on average 30% lower (more beneficial) than those from trials with adequate methodology (combined ratio of odds ratios 0.70, 95% confidence interval 0.62 to 0.80). The inadequate generation of allocation sequences was assessed in three studies only and was not consistently associated with treatment effects, although an effect was evident in the study from Denmark (figure). Interestingly, when only trials with adequate concealment of allocation were analysed in Schulz et al’s study, those with an inadequate generation of allocation sequences did yield inflated treatment effects. This indicates that if assignments are predictable some deciphering can occur, even with adequate concealment. On the other hand, the generation of unbiased sequences is probably irrelevant if the sequences are not concealed from those involved in the enrolment of patients.

Results for double blinding were more heterogeneous: the two larger studies found that estimates were on average moderately biased in open trials, whereas one of the two smaller studies showed no effect, and the other showed substantial bias associated with lack of double blinding (figure). To some extent the importance of blinding depends on the outcomes assessed. In some situations—for example, when exam-
Education and debate

Metaanalysis of four empirical studies relating key aspects of methodological quality of controlled trials to their effect estimates. Metaanalysis was by random effects model. Size of squares is proportional to inverse of variance of estimate.

**Generation of allocation sequence** (inadequate or unclear versus adequate)
- Schulz 1995: 0.95 (0.81 to 1.12)
- Moher 1998: 0.89 (0.67 to 1.20)
- Kjærsgard 2000: 0.49 (0.30 to 0.81)
- Combined: 0.81 (0.60 to 1.09)

**Concealment of allocation** (inadequate or unclear versus adequate)
- Schulz 1995: 0.66 (0.59 to 0.73)
- Moher 1998: 0.63 (0.45 to 0.88)
- Kjærsgard 2000: 0.60 (0.31 to 1.15)
- Jüni 2000: 0.79 (0.70 to 0.89)
- Combined: 0.70 (0.62 to 0.80)

**Double blinding** (absent versus present)
- Schulz 1995: 0.83 (0.71 to 0.96)
- Moher 1998: 1.11 (0.76 to 1.63)
- Kjærsgard 2000: 0.58 (0.33 to 0.98)
- Jüni 2000: 0.88 (0.75 to 1.04)
- Combined: 0.86 (0.74 to 0.99)

**Quality of reporting**

The assessment of the methodological quality of a trial is intertwined with the quality of reporting—that is, the extent to which a report provides information about the design, conduct, and analysis of the trial. Reports often omit important methodological details. For example, only 1 of 122 randomised trials of selective serotonin reuptake inhibitors specified the method of randomisation. A widely used approach to this problem is to assume that the quality was inadequate unless the information to the contrary is provided (the “guilty until proved innocent” approach). This is often justified because faulty reporting generally reflects faulty methods. A well conducted but badly reported trial will, however, be classified as poor. An alternative approach is to explicitly assess the quality of the reporting rather than the adequacy of the methods. This is also problematic because a biased but well reported trial will receive full credit. The adoption of guidelines on the reporting of clinical trials has recently improved this situation for several journals, but deficiencies in reporting will continue to be confused with deficiencies in design, conduct, and analysis.

**Assessing trial quality**

How the quality of trials should be assessed is being debated. Quality scales combine information on several features in a single numerical value, whereas the component approach examines key dimensions individually, without calculation of a score. Moher et al reviewed the use of quality scores in systematic reviews published in medical journals and the Cochrane database of systematic reviews. Trial quality was assessed in 78 (38%) of the 204 reviews from journals, of which 20 (26%) used components and 52 (67%) used scales. By contrast, all 36 reviews from the database assessed quality, of which 33 (92%) used components and none used scales.

Scales vary considerably in dimensions covered and complexity. Many scales include items for which there is little evidence that they are related to the inter-

Dimensions of external validity

External validity relates to the applicability of the results of a study to other “populations, settings, treatment variables, and measurement variables”. External validity is a matter of judgment, which depends on the characteristics of the patients included in the trial, the setting, the treatment regimens, and the outcomes assessed (box). In recent years large meta-analyses based on data from individual patients have shown that important differences in treatment effects may exist between patient groups and settings. For example, antihypertensive treatment reduces total mortality in middle aged patients with hypertension, but this may not be the case in elderly people. The benefits of fibrinolytic treatment in suspected acute myocardial infarction has been shown to decrease linearly with the delay between the start of symptoms and the initiation of treatment. In trials of cholesterol lowering drugs the benefits of a reduction in non-fatal myocardial infarction and mortality due to coronary heart disease depends on the reduction in total cholesterol concentration and the duration of follow up. At the very least, therefore, assessment of a trial’s applicability requires adequate information about the characteristics of the participants.
nal validity of a trial. For example, a widely used instrument includes items related to the presentation of data and the organisation of the trial. Unsurprisingly, different scales can lead to discordant results. This was shown in a study in which 25 different scales were used to assess 17 trials comparing low molecular weight heparin with standard heparin for thromboprophylaxis. With some scales, the relative risks of the “high quality” trials were close to unity and not statistically significant, indicating that low molecular weight heparin was not superior to standard heparin, whereas the “low quality” trials assessed by these scales showed better protection with the low molecular weight heparin. With other scales the opposite was the case: high quality trials indicated that low molecular weight heparin was superior to standard heparin, whereas low quality trials found no significant difference.

When the association of effect estimates with quality scores is examined, interpretation of results is difficult. In the absence of an association there are three possible explanations: there is no association with any of the components; there are associations with one or several components, but these components have so little weight that the effects are lost in the summary score; or there are associations with two or more components, but these cancel out so that no association is found with the overall score. On the other hand, if treatment effects do vary with quality scores then investigators will have to identify the component or components that are responsible for this association to interpret this finding.

The analysis of individual components of trial quality overcomes many of the shortcomings of composite scores. The component approach takes into account that the importance of individual quality domains, and the direction of potential biases associated with these domains, varies between the contexts in which trials are performed.

Incorporating study quality into meta-analysis

It makes intuitive sense to take into account information on the quality of studies when doing systematic reviews. One approach is to exclude trials that fail to meet some standard of quality. This may often be justified but could exclude studies that might contribute valid information. It may therefore be prudent to exclude only trials with gross deficiencies in design—for example, those that clearly failed to study comparable groups. The possible influence of study quality on effect estimates should, however, always be examined in a given set of included studies. Several approaches have been proposed for this purpose.

Quality as a weight in statistical pooling

The most radical approach is to directly incorporate information on study quality as weighting factors in the analysis. Study weights can be multiplied by quality scores, thus increasing the weight of trials deemed to be of high quality and decreasing the weight of those of low quality. A trial with a quality score of 40 out of 100 will thus get the same weight in the analysis as a trial with half the amount of information but a quality score of 80.

Weighting by quality scores is problematic for several reasons. As mentioned, the choice of the scale influences the weight of individual studies in the analysis, and the combined effect estimate and its confidence interval therefore depend on the scale. However, there is no reason why study quality should modify the precision of estimates. Poor studies are still included. Thus any bias associated with poor methodology is only reduced, not removed. Including both good and poor studies may also increase heterogeneity of estimated effects across trials and may reduce the credibility of a systematic review. The incorporation of quality scores as weights lacks statistical or empirical justification.

Sensitivity analysis

The robustness of the findings of a meta-analysis to different assumptions should always be examined in a thorough sensitivity analysis. An assessment of the influence of methodological quality should be part of this process. Simple stratified analyses and meta-regression models are useful for exploring associations between treatment effects and study characteristics. Quality summary scores or categorical data on individual components can be used for this purpose. For the reasons discussed the authors recommend that sensitivity analysis should be based on the components of study quality that are considered important in the context of a given meta-analysis. Other approaches, such as plotting effect estimates against quality scores or performing cumulative meta-analysis in order of quality, are also affected by the problems surrounding composite scales.

Conclusions

There is ample evidence that many trials are methodologically weak and increasing evidence that deficiencies translate into biased findings of systematic reviews. The assessment of the methodological quality of controlled trials and the conduct of sensitivity analyses should therefore be considered routine procedures in systematic reviews and meta-analysis. Although composite quality scales may provide a useful overall assessment when comparing populations of trials, such scales should generally not be used to identify trials of apparent low quality or high quality in a given systematic review. Rather, the relevant methodological aspects should be identified a priori and assessed individually. This should include the generation and concealment of treatment allocation,
blinding, and handling of attrition in the analysis. Other ways of investigating and dealing with bias in systematic reviews will be discussed and illustrated later in this series.37

We thank Ken Schulz and Lise Kjaergard for unpublished data and Iain Chalmers for useful comments on an earlier version of this paper.

Funding: PJ is supported by the Swiss National Science Foundation. The work on trial quality in Bristol was supported by the NHS Research and Development Programme. Competing interests: None declared.


