Published in final edited form as:
Rev Recent Clin Trials. 2009 May ; 4(2): 79–88.

A General Framework for the Evaluation of Clinical Trial Quality

Vance W. Berger[1] and **Sunny Y. Alperson**[2]

[1]*Vance W. Berger, NIH, Suite 3115, 6130 Executive Boulevard, MSC-7354, Bethesda, MD 20892-7354 U.S.A*

[2]*Sunny Y. Alperson, California State University San Marcos, Department of Nursing, CA 92096-0001*

Abstract

Flawed evaluation of clinical trial quality allows flawed trials to thrive (get funded, obtain IRB approval, get published, serve as the basis of regulatory approval, and set policy). A reasonable evaluation of clinical trial quality must recognize that any one of a large number of potential biases could by itself completely invalidate the trial results. In addition, clever new ways to distort trial results toward a favored outcome may be devised at any time. Finally, the vested financial and other interests of those conducting the experiments and publishing the reports must cast suspicion on any inadequately reported aspect of clinical trial quality. Putting these ideas together, we see that an adequate evaluation of clinical quality would need to enumerate all known biases, update this list periodically, score the trial with regard to each potential bias on a scale of 0% to 100%, offer partial credit for only that which can be substantiated, and then multiply (not add) the component scores to obtain an overall score between 0% and 100%. We will demonstrate that current evaluations fall well short of these ideals.

Keywords

Additive; Evaluation Systems; Randomization; Trial Quality

1. Introduction

In many cases, flawed or misleading evidence is worse than no evidence at all. This is because the state of ignorance resulting from a lack of evidence is recognized as a state of ignorance, whereas the state of ignorance resulting from misleading evidence is not so recognized. In addition, the existence of any clinical trials, misleading or not, effectively precludes the possibility of planning future trials to address the same questions as those addressed by the existing trials. For these reasons, misleading evidence in the form of flawed clinical trials is quite troublesome to public health. There are many contributions to the flaws routinely seen in trials, but perhaps the most important is the fact that flawed *evaluation* of trial quality allows flawed trials to thrive (get funded, obtain IRB approval, get published, serve as the basis of regulatory approval, and set policy). Hence, the assessment of validity has been identified as one of the most important steps of the peer-view process [1], and as one of the key components of systematic reviews [2].

Yet some studies rated by popular evaluations as “high quality” have been found subsequently to have major problems [3] [4]. The failure of these evaluation systems is instructive. In Section 2 we discuss some popular evaluation systems, including the Jadad score [5], the Delphi List [6], the CONSORT statement [7], and the Cochrane Back Review Group criteria [8]. In Section 3 we describe the shortcomings of each of these evaluation systems. In Section 4 we enumerate the characteristics required of a more suitable evaluation system. Specifically, we note that a reasonable evaluation of clinical trial quality would recognize that any one of a large number of potential biases could by itself completely invalidate the trial results. Hence, the evaluation would need to be comprehensive, and address each potential bias. Moreover, it would need to evaluate each bias effectively which, as we shall see, is no trivial step. In addition, clever new ways (intentionally or unintentionally) to distort trial results toward a favored outcome may be devised at any time, so the evaluation would need to be updated periodically. Finally, the vested financial and other interests of the experimenters and authors need to cast suspicion on any aspect of clinical trial quality that is inadequately reported. Putting these ideas together, we see that an adequate evaluation of clinical quality would need to enumerate all known biases, update this list periodically, evaluate each item stringently, score the trial (from 0% to 100%) with regard to each potential bias, give credit only when there is sufficient information to justify it, and then multiply (not add) the component scores to obtain an overall quality score between 0% and 100%. In future work we intend to develop such a scale, but the scale itself is beyond the scope of this paper, which aims only to justify the need for such a comprehensive approach, and to highlight the characteristics of an effective evaluation system for the quality of clinical trials.

2. Current Trial Evaluation Systems

There are several formal systems in widespread use for evaluating trial quality, as well as some *ad hoc* ones. We will discuss some of the most popular among these in this section, and note that there is considerable, but not complete, overlap among these evaluations (see Table 1 below). This is to be expected, and of course is not in and of itself a weakness of any of the instruments.

1) Ad Hoc Methods

Rather than using a general evaluation tool, many systematic reviews develop an *ad hoc* criteria list to assess the methodological quality of the primary studies for inclusion in the review and other purposes. For example, Amori et al. [9] conducted a review in which “differences in baseline characteristics between groups, description of allocation concealment, intention-to-treat analysis, and dropout rate were used to evaluate study quality.” Many systematic reviews do not evaluate the quality of the studies at all, which is in itself a kind of *ad hoc* evaluation, because it effectively assigns a perfect quality-score to each trial. Moher et al. [10] found that trial quality was assessed in only 48% of their sample of 240 meta-analyses. In addition, information on the reproducibility of the assessments, which is a hallmark of the scientific method, was provided by only half of that 48%. Only 25% of the meta-analyses that assessed quality incorporated trial quality into their analyses.

2) Jadad Score [5]

The Jadad score [5] is often used to assess the methodological quality of controlled trials. Studies are scored according to the presence of three key methodological features of clinical trials, specifically randomization, masking, and accountability of all patients, including withdrawals. One point is added for a “yes” answer to each of the first five items, and one point is subtracted for a “yes” answer to either of the last two items, for an overall score from 0–5. Its short length and ease of use remain as attractions; the 3 feature brevity gives the least responder burden [38]. Appendix 1 displays the items used in calculating the Jadad score, along with their guidelines for assessment.

3) Delphi [6]

Verhagen et al. [6] developed a list of criteria to evaluate trials by consulting a group of 33 experts in RCT evaluation. They used the Delphi consensus method to reduce 206 items from existing criteria lists into a list of nine items. Though each item is ostensibly a measure of quality, the authors noted that they did not reach a consensus on the definition of quality. There was no mention of how to calculate a numerical score from the list, but it has been implemented in practice by counting the number of positive responses to the nine questions [11] [12]. Appendix 2 gives the final Delphi list.

4) CONSORT [7]

The Consolidated Standards of Reporting Trials (CONSORT) statement is an international effort to provide instructional guidance for authors on how to report the clinical trials. The importance of reporting goes without saying; inaccurate or vague reports of trials often compound the difficulty of evaluating overall quality. The revised version of the CONSORT statement in 2001 presents 22 comprehensive checklist items along with blank space for recording the pertaining page numbers from the trial reports. Though one can refer to this checklist as a guide for design and analysis in a research study, it does not rate the actual quality of reported items. Thus, CONSORT explicitly states that it “is not meant to be used as a quality assessment instrument” [7]. However, it is often used in practice as a *de facto* quality assessment for meta-analyses [13] [14]. The CONSORT checklist is given in Appendix 3.

5) Cochrane Collaboration [8]

Quality criteria from the Cochrane Back Review Group have been used to evaluate trial quality in various reviews [15] [16]. The Cochrane collaboration is widely believed to be a quality rating standard and has many uncritical followers among researchers and reviewers. Close examination of Table 1 reveals great overlap with the Jadad scale with the exception of two items, allocation concealment and baseline data. Appendice 4 and Appendice 5 show the Cochrane Back Review list, along with recommendations for evaluating each item.

3. Shortcomings of Current Evaluation Methods

In this section we outline flaws of the current popular methods, including 1) incompleteness, 2) failure to adequately evaluate the criteria purported to be evaluated, 3) willingness to offer partial credit for missing information, 4) failure to be periodically updated, and 5) the use of an additive (not multiplicative) scoring system that allows compensation for weaknesses. It will turn out, as we shall explain, that many of the missing elements that make these evaluations incomplete may also be cast as inadequate evaluations of existing criteria.

3.1 Incompleteness

As can be seen from Table 1 above, the scales considered can be aggregated by taking the union of the elements considered in each. This would produce a set of ten items; no single one of the scales addresses all ten of these items. Specifically, the Jadad scale rates only three of these ten dimensions, Delphi rates six of ten, the Cochrane list rates five of the ten, and the CONSORT report guide is more comprehensive than the other three systems as it rates nine of the ten. The purpose of the Delphi project seems to have been to simplify, from a pool of 206 possible aspects of trial quality, down to a final list of only nine aspects. As Albert Einstein has said, "everything should be made as simple as possible, but not simpler." Unfortunately, many of the evaluation methods are overly simplistic. We argue that a suitable evaluation must be comprehensive, removing only those elements that are truly redundant or unnecessary. The following two examples demonstrate how studies can receive quality ratings of being methodologically “perfect” trials by current systems, yet be badly (possibly fatally) flawed.

Example 1—Collier et al. [19] examined the methodology employed in systematic reviews. They compared 28 systematic reviews from the Cochrane Skin Group to 10 produced elsewhere. The authors found that the Cochrane methodology resulted in higher quality reviews of dermatology studies. This finding is not particularly surprising, given the conflicts of interests for some of the authors. Specifically, one author is the Coordinating Editor of the Cochrane Skin Group, and another is the Director of their U.S. satellite group. Even assuming their impartiality, we believe that Collier's overview was not nearly a sufficiently critical assessment, because it failed to mention that the reviews it examined did not themselves adequately assess trial quality. As an example, Gibb's review [20] received a rating of 7 out 7 using the Oxman and Guyatt Scale [2]; and one unconvincing trial contained therein by Stender et al. [21] received the highest rating possible. This study on photodynamic therapy for hand and foot warts presents serious methodological flaws. First, Stender et al. [21] randomized warts, instead of patients, apparently without addressing or correcting for the lack of independence this creates among the measurements across warts from a given patient. In addition, though the use of intent-to-treat was a major contributor to the high rating this trial received, in fact the trial did no such thing, as three randomized warts per group were excluded from the analysis, the very antithesis of the principle of intent-to-treat. Allocation concealment was also a basis for the high rating, yet in point of fact allocation concealment is questionable at best, given the flawed randomization method (permuted blocks of size two) that was used. Also, analyses such as the analysis of variance and the chi-square test were used instead of exact analyses that would not have required unreasonable distribution assumptions. Clearly, the Stender trial was far from perfect.

Example 2—Similarly, Bookman et al.'s trial in Towheed's systematic review [35] received a perfect quality score 5 out of 5 using the Jadad scale. This study [18] was purportedly a high quality trial demonstrating the effectiveness of Pennsaid on patients with osteoarthritis of the knee. But in reality, the methodology of the study displayed multiple flaws and weaknesses in design and analysis which preclude the possibility of drawing defensible conclusions. These issues include possible unmasking due to recognizable treatment side effects, with 36% of the treatment group patients experiencing dry, flaky, skin as compared with only 1% of the placebo group patients; the predictability of allocation due to inadequate block size in randomization; possible selection bias as revealed in baseline imbalances of treatment and placebo patient groups; the treatment of non-numeric Likert scale items as numeric; missing baseline data, making endpoint determinations moot; the measurement of "baseline" scores subsequent to randomization; the unplanned interim analysis increasing the sample size, with no penalty; the imputing of missing data by using the last available observation without a sensitivity analysis; the post-randomization exclusion and misrepresentation of the analysis as "intent-to-treat"; and the use of ANCOVA as the main analysis without reported verification of assumptions underlying ANCOVA. More detailed critique of this study can be found in a letter to the editor which followed publication of the systematic review [3].

The low validity of systematic reviews utilizing these popular yet incomplete quality rating instruments is disturbing, because these flaws are not isolated phenomena to these two examples provided here, but rather are representative of numerous flawed ratings of studies in systematic reviews. And those systematic reviews will mislead readers and policy-makers, who assume that the stated conclusions are warranted.

3.2 Inadequate Evaluation

Even the aspects of quality that are included in the assessment tools tend to be evaluated inadequately. Currently used systems tend to employ a list of items with binary responses, precluding the possibility of accounting for the many ways that each quality element can be tainted. Additionally, the evaluations tend to rely exclusively on self-reporting. Consider a criminal trial in which the defense lawyer asks the defendant on the witness stand: "Did you commit this crime?" An answer of "no" would hardly be sufficient to exonerate the defendant. As an example, one question that is asked is "Was the study randomized?" and possibly "Was the method of randomization appropriate?" with no real evidence required. For example, the Cochrane Back Review Group lists "sealed, opaque envelopes" as an adequate form of generating random assignment sequence (see Appendix 5). It seems that this is not even a form of generating an assignment sequence at all, good or bad. A study may claim to be randomized, but may still in fact not be [22]. In addition, even if the study was randomized, it still may have been randomized poorly. A randomization scheme that is not ideal offends more than academic standards. It also compromises the integrity of the treatment comparisons, in that it practically ensures non-comparable groups.

The details of the randomization procedure are the key to its success, and inadequate attention to these details can seriously compromise the validity of the conclusions. A lack of true randomization, even if alternation (sometimes referred to as “quasi-randomization”) is used, can lead to numerous biases and can cause a lack of masking. It is partially for this reason that we agree with Bath et al. [17] that quasi-randomization methods such as alternation are unacceptable (especially for drug trials). Furthermore, some methods of randomization are better than others, in terms of predictability [30]. The most desirable method of randomization is one in which the treatment group for each patient is truly random, and completely unpredictable. For two treatment groups, this means that for each patient, the probability of allocation to either group would be one half, even once prior allocations are known. However, this unrestricted randomization is not often used in practice because it allows for both unequal sample sizes and chronological bias [31]; other more predictable methods tend to be employed instead. For example, it is common to use permuted blocks, which, by balancing within the blocks, accomplish the goal of ensuring balance and reducing chronological bias. However, this is a Pyrrhic victory, because permuted blocks also allow for substantial prediction of future allocations, and therefore can pose a major threat to allocation concealment. This is true especially, but not exclusively, for trials with a fixed small block size, as was the case for the Bookman et al. trial in Example 2.

Consider the claim and the actual success of the masking. That is, masking means more than simply attempting to conceal treatment identities; it requires the success of this effort [3]. While masking can be evaluated, it rarely is. One way to check the success of masking is illustrated by Shen et al. [32], who examined the result of masking by surveying patients to determine if they had any knowledge of their assigned groups, and to assess the attitude of the clinician who administered treatment, the technical quality of the treatment, and the degree of friendliness. Presumably, the latter would vary across groups if there was unmasking of the physicians. Of course, it is also possible to ask physicians what they think each patient received, but in the case of subversion there would be a vested interest to not reveal what was known, so this could well be unreliable [31]. The Berger-Exner test [31] is perhaps the best way to assess the success of masking, as it is objective, and does not rely on good intentions or good recall. Instead, it examines the association between a “reverse propensity score” and objective response variables within each treatment group. Positive findings from this test and its graph and a lack of imbalances among recorded covariates indicate an unbalanced latent covariate or unobservable third-order selection bias; negative findings can rule out selection bias.

Another issue is the suitability of the statistical analysis. One must bear in mind the sharp distinction between common analyses and valid analyses. Just as popular opinion did not clothe the naked emperor, so too is it the case that the popularity of a statistical method cannot be taken as a valid substitute for its applicability and appropriateness. As Point #A7 of the Ethical Guidelines for Statistical Practice of the American Statistical Association points out, “The fact that a procedure is automated does not ensure its correctness or appropriateness”. Point #A5 adds “Use only statistical methodologies suitable to the data and to obtaining valid results”. Adequate statistical analyses do not go beyond the realities of the way in which the data were collected. For example, assuming that the data have a normal distribution, or that two groups will have equal variances, or that hazards are proportional, or that odds ratios across strata are common, will weaken the subsequent analysis, because the validity of the analysis will be predicated on the truth of the assumption(s). It is far preferable to use an exact test rather than a chi-square test or t-test.

3.3 Awarding Partial Credit for Missing Information

Lexchin [99] (page 419) notes “a subtle but unmistakable shift in thinking from the precautionary principle to risk management. Precaution means assuming that a new product is potentially harmful until proven otherwise whereas the principles of smart regulation and risk management are based on the premise that there would have to be a threat of serious or irreversible damage from a product before a regulatory body would step in.” Contributing to this most disturbing trend is the lax handling of missing information. In fact, several current trial evaluations are far too forgiving of missing information. For example, the Jadad score awards one point if a trial is randomized and an extra point if the method of randomization was mentioned and appropriate. If the report mentioned the method, but the method was inappropriate, then a point is deducted. The problem is that if the method is not specified, then the evaluator is to assume the method was rigorous, and no point is deducted. This convention creates a disincentive for investigators to provide relevant information. Imagine if one could earn full credit on an exam by leaving some (or even all) answers blank, or if one could obtain a passport or a driver’s license or gainful employment by filling in only some, but not all, of the required information.

Some have argued that failure to describe methodology is a problem in reporting, not methodology. That may be true, but reporting is important insofar as it allows the reader to verify key aspects of internal validity. Soares et al. [27] recommend that if there is missing information about some aspect of the trial, then the quality evaluator should read the original research protocol or contact the authors of the study. One problem with this approach is that authors can often be remiss in responding to requests for information about previous articles. Wong et al. had this experience when conducting a review of treatments for diabetic neuropathy [28]. They sent 25 letters to authors of previous studies for further information about their reports, including their methods of randomization, allocation concealment, masking, outcome measures, and information about dropouts. Of the 25 authors they contacted, only two of them replied. Along these same lines, Lexchin [99] (page 421) found that “faculty members with industry support were almost twice as likely as those without such support (11.1% versus 5.8%) to report that they had refused requests from other academic scientists to share research results or biomaterials.” Thus, for a study to be credible, it must meet the burden of proof by clearly articulating the sound methodology which was employed. Ambiguity must not be rewarded.

3.4 Static List

The current methods of assessment have insufficient provision for updates. Just as antivirus software requires frequent updates to keep up with the latest threats, an evaluation system needs to be continually updated to reflect new possible violations of trial integrity. As a case in point, some flawed methods of randomization such as inadequate block sizes are still being used, despite the publication of methods to avoid them. The Jadad score was compiled before many problems were so recognized, so it is powerless to protect against them. Despite this obvious omission, the Jadad score continues to be used for evaluating quality of trials. Its low responder burden remains an attraction, with the missing elements casually forgiven or overlooked. This thinking is flawed. We now understand that randomization without allocation concealment can lead to selection bias, and that it can completely invalidate any claim of causality in a trial. As we study trial designs, we are learning more about potential biases, and these should be reflected in the assessment tools through a continuing, current, and detailed reassessment of methods employed. CONSORT is an exception, in that it is updated periodically. However, one author (VWB) has had no luck at all in having important points added in to CONSORT, and so the conclusion seems to be that all revisions must come from within.

3.5 Additive, Binary Scoring Systems

Most evaluation systems score each (binary) criterion individually (0 or 1) and then sum the scores. This method seems reasonable, but it is actually flawed for at least two reasons. First, as discussed in Section 3.2, aspects of trial quality tend not to be binary, and should be rated on a scale of 0 to 1, allowing for the possibility of intermediate values. Second, the scores should not be added, even if they are graded continuously, as we shall illustrate with some analogies. First, imagine rating the overall health of an individual by rating the health of each organ, and then summing these. Now imagine an individual having good musculature, vision, hearing, and so on, but having had a fatal heart attack. We would score each organ as 100%, except for the heart, which we would score as 0%. Excellence in all other organs does not compensate for the one zero, yet by the common scoring system, the health of this deceased individual would appear to be excellent. Similarly, a student who does well in all subjects but one, and fails that one, will probably not be allowed to advance to the next grade. Another analogy might be repeatedly doubling a bet, say hoping to get to $1000 from $2 (keeping in mind that 2^10=1024). One would have to win each of the nine bets. Winning the first eight and losing the ninth would result in a total loss; there is no partial credit.

Now suppose that one of the items on a trial quality checklist receives a score of zero, but the others all have positive scores, maybe even perfect. The mean score will look quite good, depending on how many items there are. But this is specious, because just one bias has the potential to completely invalidate the trial results. Suppose, for example, that a masked trial did a perfect job in randomization, but it failed to account for withdrawals and dropouts. The trial would get a Jadad score of four points (80%), but can one say that it is a high-quality trial? What if half of the trial subjects were withdrawals or dropouts? This example shows how partial credit can fail to tell whether a trial was reliable. If we had instead multiplied the zero, then the resulting score would be zero, clearly showing the trial to be unreliable. The simple rule that anything plus zero equals itself, but anything multiplied by zero equals zero, gives us a way to compile itemized scores in a way that demonstrates the actual quality of the trial. Adopting this multiplicative system encourages researchers to adopt research methodologies with higher internal validity, thus improving the quality of information available to clinicians, policy-makers and the general public.

4. Detailed Criteria for Trial Quality

To target the problems inherent in the current quality instruments, we recommend additions to the domains of evaluation of trial quality, as listed in Table 2 below. First, the quality of randomization needs to be considered, beyond just the existence of randomization, whether just claimed or actual. It is abundantly clear that some methods are better than others to prevent prediction of future allocations based on past ones [31], so better methods should earn more credit than poor ones. This entails more precise reporting of specific randomization methods used. Likewise, masking and allocation concealment require far more specification of procedure than is usually provided, and certainly more than a vague and unverifiable statement that these elements were satisfied. An assessment of the success of masking and allocation concealment should include attention to the potential unmasking of researchers, patients, and outcome assessors as appropriate. In addition, no evaluation of the success of masking and allocation concealment is complete without the Berger-Exner test [31].

Withdrawals and dropouts should also be handled properly, and this is true even when they occur *prior* to randomization. Run-in phases, especially if they are used for patient selection, can easily compromise validity by at least two distinct mechanisms. First, offering patients the active treatment during the run-in phase (or any other time, for that matter) can induce a dependency that can easily be mistaken for a treatment effect when the group subsequently not randomized to receive this active treatment does not fare as well as the group that is. This is true whether or not the run-in phase is used for patient selection. Second, patient selection deriving from run-in phases tends to be favorable to the active treatment. That is, if the active treatment is administered during the run-in phases, then generally the patients who go on to get randomized are those who respond well in some way, or at least do not have serious adverse events, or stop taking the medication. If, on the other hand, it is the control (including placebo) given during the run-in phase, then generally the patients who go on to get randomized are those who do not respond well. Either way, the results are based on only those patients who are most favorable to the active treatment. Moreover, this group cannot even be defined without first treating the patients, so it is not a group that can be reproduced in clinical practice. The result is an overly favorable impression of the active treatment, and a lack of validity.

Of course, withdrawals and drop-outs occur also during the randomized phase, and these must be handled properly to ensure validity, and to avoid biases. Generally, some form of sensitivity analysis will be necessary. The intent-to-treat analysis aims to prevent attrition bias from disruption of baseline comparability of randomization. Thus, it includes a full compilation of all patients randomized, whether they dropped out or not. The post-randomization patient exclusion in Bookman et al.'s study [18] is a violation of the intent-to-treat analysis. The two major threats to the intent-to-treat analysis are 1) its simply not being used and 2) the claim that it is used when in fact it is not. Clearly, the second offense is the more serious one, as it involves deliberate misleading of the readers. Generally, when this is done, randomized patients who do not receive adequate treatment and/or do not contribute adequate data are excluded.

Any baseline data that are collected need to be collected before randomization, not after, because otherwise they may be influenced by treatments [999]. In the Bookman trial [18], several investigators did not collect baseline data from patients before randomization, so that 77 baseline values for a physical function and stiffness measurement were imputed from the day one, post-randomization values, and 29 baseline values for a pain measurement were similarly imputed.

Several issues arise under the umbrella heading of endpoints, including the need to specify these endpoints in the protocol and then stick with the endpoints so specified, the need for clinically meaningful (as opposed to surrogate) endpoints, and the need for maximally informative (as opposed to artificially dichotomized) endpoints. Pre-specified analyses need to be articulated before conducting the trial, including reasonable endpoints. While surrogate endpoints are often used to substitute for another endpoint which might take too long or be impractical to obtain, and though they may be highly correlated with the primary outcome variable of interest, they still cannot serve as a valid replacement. For instance, cholesterol levels have been used as a surrogate for prevention of death and heart disease in cardiovascular studies. However, individuals can have high cholesterol levels and no heart disease, and individuals with normal cholesterol levels can develop fatal heart disease. Sometimes combined endpoints are employed, checking for the occurrence of any of a number of outcomes which are clinically meaningful. For instance, one can look for severe chest pain, myocardial infarction, or death as a combined endpoint. But this treats the component endpoints as interchangeable; clearly, they are not. While post hoc analyses from subgroups here are tempting, they should be used only to suggest directions for future studies, and not to draw conclusions.

Stopping rules specifically state when a trial is to end. Sometimes, treatments that have an unplanned interim analysis do not stop the treatment at the specified time, but instead select a different stopping point. If there is an interim analysis, a penalty must be applied to preserve the Type I error rate. This was not done in the Bookman et al. study [18].

While there are many different statistical methods used to analyze the outcomes of clinical trials, the chosen analysis should be appropriate for the type of data collected. It is not enough to simply state that a particular statistical method was used; the reason that specific method is more appropriate should also be clarified. The use of a simplistic flowchart of scales of data and an application of an implied inferential test is not sufficient. For instance, many parametric tests demand special consideration before they can be applied. ANCOVA requires normality of residuals, equal variances, linearity, and independence, and it is highly unlikely that these assumptions can all be met. When these assumptions do not hold, the results may not be valid. The existence of exact analyses renders the use of approximate parametric analysis a weakness, for which credit should be withheld. Finally, multiplicity must be addressed. Many clinical trials use multiple endpoints to assess the therapeutic effects and any toxicity of a treatment. Informative endpoints fuse the set of selective endpoints in order to eliminate a series of binary endpoints. Guidelines are available [25]. Whenever there is multiple testing, *p*-values should be adjusted for multiple hypotheses testing.

5. Conclusion and Recommendations

In light of the pitfalls of trial quality assessment, the current assessment tools seem about as useful as a porous dam, or a dam that blocks the flow of only half a river. Some have argued that we should abandon this practice altogether, that is, there are too many barriers and limitations for the evidence to be trustworthy and credible. For clinicians, efforts to be more evidence-based in their practice result in contradictory conclusions, and they find it easier to resort to intuition or past experience [33]. Others have argued about the epistemological base of “evidence.” French [34] found that the notion of “evidence-based practice” is mostly subjective and it can be construed as nothing more than a “novelty effect in basically political scenarios.”

Our position is neither of these extremes. We believe that evidence-based practice since its inception has traveled a long way, and it has a considerable distance to go. We parallel Churchill, who said, “Democracy is the worst form of government except all those other forms that have been tried from time to time.” Having witnessed the replacement of expert narratives with systematic reviews, now we wait for a system that will reflect sound reason and transparent science to evaluate the quality of trials for the best healthcare practice possible. An important point to remember in this endeavor, as Feigenbaum and Levy [29] point out, is that biases are hardly haphazard. They tend to go in the intended direction, bearing in mind the strong interests in particular outcomes on the part of the experimenters. In his defense of Bookman et al.’s study [18], Towheed states that the study is reasonably robust, and research studies are seldom methodological masterpieces of perfection and rigor [35]. This statement reinforces a sloppy status quo mentality, overlooking flaws in studies and quality evaluation, and therefore is detrimental to the advancement of better evidence-based practice.

By finding ways to better evaluate biases, subversions and various loopholes in the trial quality, we are perhaps one step closer to providing unbiased information to patients, clinicians, researchers, and policy-makers. We wish to emphasize that it is critical to remember that bad or good assessments of trial quality encourage bad or good studies to continue. Summarizing the evaluation of the current methods, we recommend that the following be incorporated into the evaluation systems to improve its quality:

1. Be comprehensive.
2. Evaluate each criterion effectively.
3. Give points only when there is information to justify the points.
4. Update the assessment tool regularly.

5. Multiply all scores instead of adding them.

References

1. Kassirer JP, Campion EW. Peer review: crude and understudied, but indispensable. Journal of the American Medical Association 1994;272:96–97. [PubMed: 8015140]

2. Oxman AD, Guyatt GH. Guidelines for reading literature reviews. Can Med Assoc J 1988;138:697–703. [PubMed: 3355948]

3. Berger VW. Is the Jadad Score the Proper Evaluation of Trials. Journal of Rheumatology 2006;33(8):1710. [PubMed: 16881132]

4. Gee E, Berger VW. On confusing prima-facie validity with true validity. British Journal of Dermatology 2007;157(2):425–426. [PubMed: 17596152]

5. Jadad, Alejandro R., MD, Dphil; Moore, R Andrew, DPhil; Carroll, Dawn, RGN; Jenkinson, Crispin, DPhil; Reynolds, D John M., DPhil; Gavaghan, DavidJ, DPhil; Henry, J.; McQuay, DM. Assessing the Quality of Reports of Randomized Clinical Trials: Is Blinding Necessary? Controlled Clinical Trials 1996;17:1–12. [PubMed: 8721797]

6. Verhagen, Arianne P.; de Vet, Henrica CW.; de Bie, Robert A.; Kessels, Alphons GH.; Boers, Maarten; Bouter, Lex M.; Knipschild, Paul G. The Delphi List: A Criterial List for Quality Assessment of Randomized Clinical Trials for Conducting Systematic Reviews Developed by Delphi Consensus. Journal of Clinical Epidemiology 1998;Vol. 51(No 12):1234–1251.

7. Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, Gøtzsche PC, Lang T. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Ann Intern Med 2001;134(8):663–694. [PubMed: 11304107]

8. van Tulder MW, et al. Updated Method Guidelines for Systematic Reviews in the Cochrane Collaboration Back Review Group. Spine 2003 June 15;28(12):1290–1299. [PubMed: 12811274]

9. Amori RE, Lau J, Pittas AG. Efficacy and safety of incretin therapy in type 2 diabetes: systematic review and meta-analysis. Journal of the American Medical Association 2007 Jul 11;298(2):194–206. [PubMed: 17622601]

10. Moher D, Cook DJ, Jadad AR, Tugwell P, Moher M, Jones A, Pham B, Klassen TP. Assessing the quality of reports of randomized trials: implications for the conduct of meta-analyses. Health Technology Assessment 1999;3(12):i–v. 1–98. [PubMed: 10374081]

11. Damen L, Bruijn JK, Verhagen AP, Berger MY, Passchier J, Koes BW. Symptomatic Treatment of Migraine in Children: A Systematic Review of Medication Trials. Pediatrics 2005;116(2):e295–e302. [PubMed: 16061583]

12. Mozaffarian, Dariush, MD, MPH; Geelen, Anouk, PhD; Brouwer, Ingeborg A., PhD, et al. Effect of fish oil on heart rate in humans: a meta-analysis of randomized controlled trials. Circulation 2005 Sep 27;112(13):1945–1952. [PubMed: 16172267]

13. Zabalegui A, Sanchez S, Sanchez PD, Juando C. Nursing and cancer support groups. Journal of Advanced Nursing 2005 Aug;51(4):369–381. [PubMed: 16086806]

14. Bize R, Burnand B, Mueller Y, Cornuz J. Effectiveness of biomedical risk assessment as an aid for smoking cessation: a systematic review. Tobacco Control 2007 Jun;16(3):151–156. [PubMed: 17565124]

15. Galaal K, Deane K, Sangal S, Lopes A. Interventions for reducing anxiety in women undergoing colposcopy. Cochrane Database of Systematic Reviews 2007 Jul 18;(3):CD006013.

16. Piazzini DB, Aprile I, Ferrara PE, Bertolini C, Tonali P, Maggi L, Rabini A, Piantelli S, Padua L. A systematic review of conservative treatment of carpal tunnel syndrome. Clinical Rehabilitation 2007 Apr;21(4):299–314. [PubMed: 17613571]

17. Bath FJ, Owen VE, Bath PMW. Quality of full and final publications reporting acute stroke trials. Stroke 1998;29(10):2203–2210. [PubMed: 9756604]

18. Bookman AM, Williams KSA, Shainhouse JZ. Effect of a topical diclofenac solution for relieving symptoms of primary osteoarthritis of the knee: a randomized controlled trial. CMAJ 2004;171:333–338. [PubMed: 15313991]

19. Collier A, Heilig L, Schilling L, Williams V, Dellavalle Rp. Cochrane Skin Group systematic reviews are more methodologically rigorous than other systematic reviews in dermatology. British Journal of Dermatology 2006;155(6):1230–1235. [PubMed: 17107394]

20. Gibbs S, Harvey I, Sterling J, Stark R. Local treatments for cutaneous warts: systematic review. BJM 2000;325(7362):461.

21. Stender IM, Na R, Fogh H, Gluud C, Wulf HC. Photodynamic therapy with 5-aninolaevulinic acid or placebo for recalcitrant foot and hand warts: randomized double-blind trial. Lancet 2000;355 (9208):963–966. [PubMed: 10768434]

22. Berger VW, Bears JD. When can a clinical trial be called 'randomized'? Vaccine 2003;21:468–472. [PubMed: 12531645]

23. Berger VW, Exner DV. Detecting selection bias in randomized clinical trials. Control Clin Trials 1999 Aug;20(4):319–327. [PubMed: 10440559]

24. Berger VW. Pros and Cons of Permutation Tests in Clinical Trials. Statistics in Medicine 2000;19:1319–1328. [PubMed: 10814980]

25. Berger VW. Improving the information content of categorical clinical trial endpoints. Control Clin Trials 2002 Oct;23(5):502–514. [PubMed: 12392864]

26. Boutron I, Estellat C, Ravaud P. A review of blinding in randomized controlled trials found results inconsistent and questionable. Journal of Clinical Epidemiology 2005 Dec;58(12):1220–1226. [PubMed: 16291465]

27. Soares HP, Daniels S, Kumar A, Clarke M, Scott C, Swan S, Djulbegovic B. Bad reporting does not mean bad methods for randomised trials: observational study of randomised controlled trials performed by the Radiation Therapy Oncology Group. BMJ 2004 Jan 3;328(7430):22–24. [PubMed: 14703540]

28. Wong MC, Chung JW, Wong TK. Effects of treatments for symptoms of painful diabetic neuropathy: systematic review. BMJ 2007 Jul;335(7610):87. [PubMed: 17562735]

29. Feigenbaum, Susan; Levy, DavidM. The Technological Obsolescence of Scientific Fraud. Rationality and Society 1996;8(3):261–276.

30. Friedman, LM.; Furberg, CD.; DeMets, DL. Fundamentals of Clinical Trials. Vol. 3rd edition. New York: Springer-Verlag; 1998.

31. Berger, VW. Selection bias and covariate imbalances in randomized clinical trials. Chrichester John Wiley & Sons; 2005.

32. Shen J, Wenger N, Glaspy J, Hays RD, Albert PS, Choi C, OMD, Shekelle PG. Electroacupuncture for control of Myeloablative Chemotherapy-Induced Emesis A Randomized Controlled Trial. JAMA 2000;284:2755–2761. [PubMed: 11105182]

33. Tracy CS, Dantas GC, Upshur RE. Evidence-based medicine in primary care: qualitative study of family physicians. BMC Family Practice 2003;4:6. [PubMed: 12740025]

34. French P. What is the evidence on evidence-based nursing? An epistemological concern. Journal of Advanced Nursing 2002;37(3):250–257. [PubMed: 11851795]

35. Towheed TE. Pennsaid therapy for osteoarthritis of the knee: a systematic review and metaanalysis of randomized controlled trials. J Rheumatol 2006;33(8):567–573. [PubMed: 16511925]

36. Larzelere R, Kuhn B, Johnson B. The Intervention Selection Bias: An Underrecognized Confound in Intervention Research. Psychological Bulletin 2004;130(2):289–303. [PubMed: 14979773]

37. Khan KS, Riet G, Popay J, Nixon J, Kleijnen J. Study quality assessment (phase 5): conducting the review (stage 2). Undertaking systematic reviews of research on effectiveness. CRD Report 2001;4:1–20.

38. West, S.; King, V.; Carey, TS., et al. Rockville, MD: Agency for Healthcare Research and Quality; 2002. Systems to rate the strength of scientific evidence. Evidence Report/Technology Assessment No. 47 (Prepared by the Research Triangle Institute-University of North Carolina Evidence-based Practice Center under Contract No. 290-97-0011). AHRQ Publication No. 02-E016

39. Moye, LA. Statistical Reasoning in Medicine. Vol. 2nd edition. New York: Springer; 2006.

99. Lexchin J. The secret things belong unto the Lord our God: Secrecy in the pharmaceutical arena. 2007;26:417–430.

999. Berger VW. Valid Adjustment for Binary Covariates of Randomized Binary Comparisons. Biometrical Journal 2004;46(5):589–594.

Table 1

Dimensions of Trial Quality Measured by Assessment Tools

	Jadad	Delphi*	CONSORT*	Cochrane
Randomization	J1, J2, J6	D1a	C1, C8, C10	A
Masking	J3, J4, J7	D4, D5, D6	C11	D, E, F
Allocation Concealment		D1b	C9	B
Handling of Withdrawals and Dropouts	J5	D8	C13, C16	H, I, K
Measures of Variability		D7		
Pre-specified Analyses			C6	
Stopping rules			C7	
Statistical methods			C12, C17	
Baseline data		D2	C15	C
Address Multiplicity			C18, C20	

*We only include those items of Delphi and CONSORT that pertain to internal validity.

Table 2

Recommended Additions and Changes to Evaluation of Trial Quality

Randomization	Quality of randomization
Masking	Proper evaluation of success
Allocation Concealment	Proper evaluation of success
Withdrawals and Dropouts	Run-in selection bias Proper ITT and sensitivity analyses
Baseline data	Measured before randomization
Endpoints	Maximally informative, clinical (not surrogate) endpoints, pre-specified
Stopping rules	Account for any interim analysis
Statistical methods	Minimal assumptions, adhere to the scale of measurement, pre-specified
Measures of Variability	Necessary data table presented
Address Multiplicity	Adjust p-values for multiple testing

Appendix 1

Jadad Score Calculation

Item	Score
Was the study described as randomized (this includes words such as randomly, random, and randomization)?	0/1
Was the method used to generate the sequence of randomization described and appropriate (table of random numbers, computer-generated, etc)?	0/1
Was the study described as double blind?	0/1
Was the method of double blinding described and appropriate (identical placebo, active placebo, dummy, etc)?	0/1
Was there a description of withdrawals and dropouts?	0/1
Deduct one point if the method used to generate the sequence of randomization was described and it was inappropriate (patients were allocated alternately, or according to date of birth, hospital number, etc).	0/−1
Deduct one point if the study was described as double blind but the method of blinding was inappropriate (e.g., comparison of tablet vs. injection with no double dummy).	0/−1
Guidelines for Assessment	
Randomization A method to generate the sequence of randomization will be regarded as appropriate if it allowed each study participant to have the same chance of receiving each intervention and the investigators could not predict which treatment was next. Methods of allocation using date of birth, date of admission, hospital numbers, or alternation should not be regarded as appropriate.	
Double blinding A study must be regarded as double blind if the word “double blind” is used. The method will be regarded as appropriate if it is stated that neither the person doing the assessments nor the study participant could identify the intervention being assessed, or if in the absence of such a statement the use of active placebos, identical placebos, or dummies is mentioned.	
Withdrawals and dropouts Participants who were included in the study but did not complete the observation period or who were not included in the analysis must be described. The number and the reasons for withdrawal in each group must be stated. If there were no withdrawals, it should be stated in the article. If there is no statement on withdrawals, this item must be given no points.	

Appendix 2

Final Delphi List after three Delphi rounds [6]

Item	Answer
1. Treatment allocation a) Was a method of randomization performed?	Yes/No/Don't know
b) Was the treatment allocation concealed?	Yes/No/Don't know
2. Were the groups similar at baseline regarding the most important prognostic indicators?	Yes/No/Don't know
3. Were the eligibility criteria specified?	Yes/No/Don't know
4. Was the outcome assessor blinded?	Yes/No/Don't know
5. Was the care provider blinded?	Yes/No/Don't know
6. Was the patient blinded?	Yes/No/Don't know
7. Were point estimates and measures of variability presented for the primary outcome measures?	Yes/No/Don't know
8. Did the analysis include an intention-to-treat analysis?	Yes/No/Don't know

Appendix 3

The Revised CONSORT Statement [9]

Checklist of items to include when reporting a randomized trial BMD Appendix 3: The Revised CONSORT Statement [9]

Checklist of items to include when reporting a randomized trial

PAPER SECTION And topic	Item	Description	Reported on page #
TITLE & ABSTRACT	1	How participants were allocated to interventions (*e.g.*, "random allocation", "randomized", or "randomly assigned").	
INTRODUCTION Background	2	Scientific background and explanation of rationale.	
METHODS Participants	3	Eligibility criteria for participants and the settings and locations where the data were collected.	
Interventions	4	Precise details of the interventions intended for each group and how and when they were actually administered.	
Objectives	5	Specific objectives and hypotheses.	
Outcomes	6	Clearly defined primary and secondary outcome measures and, when applicable, any methods used to enhance the quality of measurements (*e.g.*, multiple observations, training of assessors).	
Sample size	7	How sample size was determined and, when applicable, explanation of any interim analyses and stopping rules.	
Randomization -- Sequence generation	8	Method used to generate the random allocation sequence, including details of any restriction (*e.g.*, blocking, stratification).	
Randomization -- Allocation concealment	9	Method used to implement the random allocation sequence (*e.g.*, numbered containers or central telephone), clarifying whether the sequence was concealed until interventions were assigned.	
Randomization -- Implementation	10	Who generated the allocation sequence, who enrolled participants, and who assigned participants to their groups.	
Blinding (masking)	11	Whether or not participants, those administering the interventions, and those assessing the outcomes were blinded to group assignment. If done, how the success of blinding was evaluated.	
Statistical methods	12	Statistical methods used to compare groups for primary outcome (s); Methods for additional analyses, such as subgroup analyses and adjusted analyses.	
RESULTS Participant flow	13	Flow of participants through each stage (a diagram is strongly recommended). Specifically, for each group report the numbers of participants randomly assigned, receiving intended treatment, completing the study protocol, and analyzed for the primary outcome. Describe protocol deviations from study as planned, together with reasons.	
Recruitment	14	Dates defining the periods of recruitment and follow-up.	
Baseline data	15	Baseline demographic and clinical characteristics of each group.	
Numbers analyzed	16	Number of participants (denominator) in each group included in each analysis and whether the analysis was by "intention-to-treat". State the results in absolute numbers when feasible (*e.g.*, 10/20, not 50%).	
Outcomes and estimation	17	For each primary and secondary outcome, a summary of results for each group, and the estimated effect size and its precision (*e.g.*, 95% confidence interval).	
Ancillary analyses	18	Address multiplicity by reporting any other analyses performed, including subgroup analyses and adjusted analyses, indicating those pre-specified and those exploratory.	
Adverse events	19	All important adverse events or side effects in each intervention group.	
DISCUSSION	20	Interpretation of the results, taking into account study	
Interpretation		hypotheses, sources of potential bias or imprecision and the dangers associated with multiplicity of analyses and outcomes.	

PAPER SECTION And topic	Item	Description	Reported on page #
Generalizability	21	Generalizability (external validity) of the trial findings.	
Overall evidence	22	General interpretation of the results in the context of current evidence.	

Appendix 4

Cochrane Back Review Group Criteria List for Methodological Quality Assessment

	Item	Answer
A	Was the method of randomization adequate?	Yes/No/Don't Know
B	Was the treatment allocation concealed?	Yes/No/Don't Know
C	Were the groups similar at baseline regarding the most important	Yes/No/Don't Know
D	prognostic indicators?	Yes/No/Don't Know
E	Was the patient blinded to the intervention?	Yes/No/Don't Know
F	Was the care provider blinded to the intervention?	Yes/No/Don't Know
G	Was the outcome assessor blinded to the intervention?	Yes/No/Don't Know
H	Were cointerventions avoided or similar?	Yes/No/Don't Know
I	Was the compliance acceptable in all groups?	Yes/No/Don't Know
J	Was the drop-out rate described and acceptable?	Yes/No/Don't Know
K	Was the timing of the outcome assessment in all groups similar?	Yes/No/Don't Know

Appendix 5

Operationalization of the Cochrane Back Review Group Criteria List

A	A random (unpredictable) assignment sequence. Examples of adequate methods are computer generated random number table and use of sealed opaque envelopes. Methods of allocation using date of birth, date of admission, hospital numbers, or alternation should not be regarded as appropriate.
B	Assignment generated by an independent person not responsible for determining the eligibility of the patients. This person has no information about the persons included in the trial and has no influence on the assignment sequence or on the decision about eligibility of the patient.
C	In order to receive a "yes," groups have to be similar at baseline regarding demographic factors, duration and severity of complaints, percentage of patients with neurologic complaints, and value of main outcome measure(s).
D	The reviewer determines if enough information about the blinding is given in order to score a "yes."
E	The reviewer determines if enough information about the blinding is given in order to score a "yes."
F	The reviewer determines if enough information about the blinding is given in order to score a "yes."
G	Cointerventions should either be avoided in the trial design or similar between the index and control groups.
H	The reviewer determines if the compliance to the interventions is acceptable, based on the reported intensity, duration, number and frequency of sessions for both the index intervention and control intervention(s).
I	The number of participants who were included in the study but did not complete the observation period or were not included in the analysis must be described and reasons given. If the percentage of withdrawals and drop-outs does not exceed 20% for short-term follow-up and 30% for long-term follow-up and does not lead to substantial bias a "yes" is scored. (N.B. these percentages are arbitrary, not supported by literature).
J	Timing of outcome assessment should be identical for all intervention groups and for all important outcome assessments.
K	All randomized patients are reported/analyzed in the group they were allocated to by randomization for the most important moments of effect measurement (minus missing values) irrespective of noncompliance and cointerventions.