Perspective

CLINICAL INVESTIGATION

Beyond Jadad: some essential features in trial quality

Clin. Invest. (2013) 3(12), 1119–1126

Randomized trials are considered the gold standard for design in medical research. However, poor quality trials often escape scrutiny cloaked under the umbrella term 'randomized trials', leading to misinformation and bringing tangible harm to patients who rely on only valid research evidence. To enhance both trial quality and the quality of evidence, we need to recognize potential flaws that can invalidate entire trials. This perspective will discuss some key issues in clinical trials including α preservation, surrogate and composite end points and spin reporting. In addition, current quality rating scales and a reporting guide will be examined for their capacity to evaluate and guide trial quality on the features discussed. Finally, the quality evaluating tools will be summarized and a future perspective on trial quality provided.

Keywords: α preservation • composite end points • exact test • Jadad • permutation tests • quality rating scales • spin reporting • surrogate end points • trial quality

Sunny Y Alperson[1] & Vance W Berger*[2]

[1]University of Maryland University College, 3501 University Boulevard E, Adelphi, MD 20783, USA
[2]National Cancer Institute & University of Maryland Baltimore County, Biometry Research Group, National Cancer Institute, 9609 Medical Center Drive, Rockville, MD 20850, USA
*Author for correspondence:
Tel.: +1 240 276 7142
Fax: +1 301 402 0816
E-mail: vb78c@nih.gov

Even in placebo-controlled, double-blind, randomized clinical trials, internal validity is not assured. Despite efforts to produce better evidence, many potential flaws in study design, execution, analysis and reporting curtail the quality of trial evidence [1–4]. Our aim is to review a few relevant issues in trial quality and examine the current tools that evaluate and guide clinical trials. We discuss the fundamental problem related to α preservation with the assumption of normality in clinical trials, and we promote replacing parametric tests with exact tests whenever possible. We also discuss end points including surrogate and clinical, issues with composite end points (CEPs), and the current status of spin reporting. For each of these issues we take a look at two widely used tools, Jadad and Chalmers *et al.*'s quality rating scales, and a popular reporting guide, CONSORT. Although CONSORT is not a quality rating checklist *per se*, it is often used as a *de facto* quality check list guide for the researchers [2,3]. For this reason, we included CONSORT with the two quality rating scales to examine their capacity to evaluate or guide trial quality. While we are aware of other viewpoints, we provide our perspectives for future developments in trial quality.

α preservation: permutation test, not parametric

It is an indisputable fact that no variable encountered in a clinical trial cannot truly be normally distributed, notwithstanding misguided efforts to argue to the contrary [5,6]. For example, is there a good rationale to consider distributions of blood pressure or genome-wide association as normal? Perhaps the data appear to be unimodal, or symmetric, or bell shaped, or some combination thereof. Does this establish normality? It is quite easy to state that a variable is normally distributed in the current state of research, and under this assumption, it seems to suggest that there is a commensurate simplicity in establishing normality; however, in reality, it belies the true complexity of demonstrating normality [7]. For a variable to be normally

FUTURE SCIENCE part of fsg

distributed, it would need to have a precise and specific positive probability for literally every subset of the real line. There would need to be a positive probability of values above 10 raised to the 50th power, for example. Beyond this, all the probabilities need to be consistent with each other. Of course, this is an infinite number of requirements. Aside from the utter impossibility of some of these probability requirements, there is also the small matter of establishing the required consistency among those that are possible. We would need to establish an infinite number of probability statements based on finite data. We see then that there are dual challenges, in that the data cannot possibly be normally distributed because of the values that would entail being possible, and even if somehow the data could be normally distributed, we could never establish that fact with finite data.

It is also an indisputable fact that parametric analyses are not valid when applied to variables that are not normally distributed. Let us recognize what 'valid' means in this context. It means preserving α, so that the actual Type I error rate is known with certainty in all cases to be no greater than the nominal one [5,8]. While some researchers have argued that validity extends to some α deviations, there is a strict definition of validity that applies without unnecessary compromise.

In a randomized trial, not only is there no normally distributed variable, but there is also no random sampling, so even if the underlying data were normally distributed, the test statistic still would not be [7,8]. The central limit theorem is a statement of limiting behavior, but not of the sampling distribution of a test statistic computed with a finite amount of data. The experiment consists of the random allocation of patients to treatment groups. This can be hypothetically repeated to construct a reference distribution for the purpose of computing a p-value which, if done correctly, is an exact permutation test.

To put the matter more directly, permutation tests are not based on assuming normality in the data or the statistical test. They permute the treatment assignments, computing a test statistic for each pseudo outcome, resulting in an exact probability statement. The conclusions are more robust because they do not rely on shaky assumptions about data distributions. Contrary to the fears of researchers that these permutation tests are too intense in computation and too conservative, in reality they are not difficult to conduct, and moreover, only they can guarantee validity and α preservation. One does not have to rely on the assumption of normality as with parametric tests in preserving α.

Some researchers may still feel strongly that they must use a parametric test. We recommend placing the exact test in the mainstream of research statistics for better analysis of evidence. For a more detailed discussion of the mathematical underpinnings and advantages of permutation tests, please see the references provided [3,5,8].

Examining current quality rating scales of clinical trials, the well known Jadad scale is silent on this key determinant of trial quality [2,9]. Though Chalmer's *et al.* call for 'appropriate statistical analyses' [10], they do not mention permutation tests or exact analyses. The CONSORT guide also gives little specific guidance on this issue [3,11].

End points: clinical & surrogate

A clinical end point directly affects patient well being. Common examples include death from any cause, death from disease, acute myocardial infarction (MI), and certain patient ratings of pain. A surrogate end point (SEP) is a substitute for a clinical end point that is expected to predict clinical benefit [12–14]. Examples are blood pressure and cholesterol level, and all sorts of markers and disease indicators that supposedly correspond with clinical end points. A significant advantage of SEPs is that they circumvent the need for large sample sizes by increasing event frequency. They can reduce time and costs of trials in finding therapeutic effect of a testing agent much quicker for the waiting patients, thereby lessening the number of patients who suffer [12–14]. In theory, then, the use of SEPs is a very good idea.

However, the dangers in relying on SEPs as valid predictors of clinical outcomes were exposed in unexpected ways by the CAST. CAST was a high-profile NIH-funded trial designed to test the hypothesis that encainide, flecainide or moricizine would suppress ventricular ectopy, ultimately improving survival in a placebo-controlled randomized trial with 1498 post-MI patients. It was a strong assumption then in the medical community that arrhythmia was surely associated with death. It turned out that the treatments were effective in suppressing ventricular ectopy, but these antiarrythmic drugs led to excess deaths at a higher rate than the placebo, which caused the abrupt ending of the trial. CAST II trials were then attempted again with a modified design and sample. This trial was also terminated prematurely because the patients in the experimental moricizine group had an excessive death rate during the first 2 weeks of exposure [15–17]. These trials provided clear lessons of a deadly potential when assuming a relationship between SEPs (decrease in arrhythmia) and the true clinical end point (mortality). Not only is it possible for a treatment to improve the surrogate and not improve the clinical end point, but it is also possible for a treatment to improve the SEP while making the clinical end point worse.

Despite the prominent role SEPs play in research methodology today, there is, as yet, no reliable set of

guidelines for their validation that offers protection from obtaining misleading evidence [18–20]. Prentice proposed a set of validation criteria for SEPs in 1989 [21], which states surrogate must be correlated with the clinical outcome and fully capture the net effect of treatment on the clinical outcome. Since then, several authors have challenged the relevance and validity of the Prentice Criteria, meaning that even satisfying these criteria may not be enough [18–20, 22–24]. One of the fundamental known difficulties in the accurate use of surrogates lies in the multiple pathways of disease process and intervention, while SEPs are more or less singular. Sometimes the intervention alters pathways in unexpected ways and does not capture the effect of intervention on the clinical outcome, as we have seen in CAST and other cardiac trials [15–17].

Demets and Califf presented two simple requirements for surrogate use. The first is that changes in surrogate must predict changes in clinical outcome; the other is that the surrogate must capture the effect of the intervention on the clinical outcome. However, the problem is that even when the SEP is highly correlated with the clinical end point, a treatment effect on the SEP does not necessarily translate into a comparable treatment effect on the clinical end point [25].

Similarly, in trials using cardiac output as a surrogate marker with intervention with inotropes and vasodilator compounds, these 'physiologically rational 'therapies showed short-term hemodynamic improvement, but they did not affect clinical outcomes in long-term survival [17,26]. If survival data had not been collected for these studies relying on surrogates, the therapeutic effect of the drugs would have been assumed to predict survival, and the study would have provided misleading evidence.

Currently, there is no way to ensure that a valid treatment comparison on the basis of the SEP will also serve as a valid treatment comparison on the basis of the clinical end point. This means that trials that use SEPs as replacements for clinical end points, as opposed to using the two in conjunction with each other, cannot demonstrate claims that we wish to establish. It is difficult to imagine the current research field of oncology or cardiology without SEPs. However, it is clear that some limitations are needed to reel in the use of SEPs. The reviewers of trial quality and evidence must remember that this is a methodological weakness that can potentially invalidate an entire trial, and take proper caution into valid evaluation of trial quality.

Upon examining quality rating tools and a guide that is used today, the Jadad scale fails to address end points all together [2,9]. Chalmers *et al.* [10] and CONSORT [11] have some questionnaire items concerning whether end points were tabulated, survival time and whether stopping dates were provided. Multiplicity issues of primary and secondary outcomes were mentioned, but there is no mention of SEPs. One can understand the omission of SEPs in the Chalmers scale, because it was published in 1981, before SEPs became common. But the revised 2010 version of CONSORT makes no mention of SEPs, though the authors undoubtedly created the guide out of concern for the quality of trials and report [3,27,28].

CEPs

A CEP is a combination of two or more end points fused into one variable, such as the occurrence of either death or MI [29–31]. CEPs have gained popularity due to both their capacity for handling multiple variables and their statistical benefits. Similar to the SEP, a CEP may reduce time and costs for a trial by curbing the need for large sample size due to the increased frequency of composite events. In addition, it can provide a measure of the net overall effect of an intervention without sharing α, and can occasionally be useful for avoiding a risk of bias due to competing risks [28–32].

Despite the multiple advantages, there are several issues to consider when using CEPs. First, the CEP often includes SEPs, so the aforementioned concerns apply. Second, the most common method of forming a CEP is to simply equate each of the component end points. For example, a typical CEP formed from death or MI would take the value 1 if either (or both) occurred and the value 0 otherwise. However, death is worse than MI, so in equating these two inherently unequal events the CEP is wasting valuable discriminating information.

According to several reports, problems arise due to a greater degree of variability among the individual components in a CEP. This can be in the frequency of events, in the effect size differences, or the degree of clinical meaningfulness to the patient [33–35]. The occurrence of a large difference among any of these can affect interpretation of the CEPs, reflecting a need for close scrutiny of trial quality [31,36–38].

For instance, a large difference in the frequency of events within the composite can seriously mislead stakeholders. Consider the well known DREAM study cited by Ferreira-Gonzles *et al.* [39,40]. This randomized controlled trial (RCT) examined the efficacy of rosiglitazone versus placebo in 5269 patients with no history of cardiovascular disease or glucose intolerance. The CEP consisted of incident diabetes (newly diagnosed diabetes mellitus) and death, and the result showed a much higher incidence of diabetes mellitus than death. The authors reported that rosiglitazone 8 mg per day, with other lifestyle change recommendations, reduces the risk of diabetes or death by 60%. It is easy to erroneously conclude that rosiglitazone is beneficial for both

components, diabetes mellitus and death; in truth, we cannot draw this conclusion about the reduction in mortality from the composite net result [31,36,37].

Another issue of heterogeneity of the components in the CEP is clinical meaningfulness; that is, how important is it to the patient? If the individual components in the composite have large variations in gradients of clinical intensity, then the proper interpretation of the CEP can become confusing. For instance, in a respiratory trial, one CEP combined all-cause mortality, intubation with mechanical ventilation and intense steroid therapy for patients with chronic obstructive lung disease. It is understandable that the investigators had reason to collect the steroid data, but patients are likely to think the steroid was inconsequential, even irrelevant compared with the variables such as intubation for mechanical ventilation and death [36,37].

A third heterogeneity issue in the CEP is effect size difference. Large variation in effect sizes among the components should give an immediate sense of caution for the accuracy of the trial interpretation [31,36,37]. Because the selection of individual components should be based on similar biological plausibility, it is natural to expect some degree of similarity in their treatment influence. CEPs with little variation in the effect size among the components can be straightforward to interpret. Ferreira-Gonzalez *et al.* rightly point out that when ramipril was tested with a composite of acute MI (AMI), cerebrovascular accident, and cardiovascular death, the relative risk was 0.7–0.8 for each component, thereby showing a similar influence of treatment on all components (HOPE investigators). However, in the Gerstein *et al.* DREAM study cited above, the hazard ratios between components of diabetes mellitus and death varied from 0.38 to 0.90, with a claim of a 60% net reduction on both components. In fact, 93% of the events were due to reduction in diabetes, with comparable low incidence of death in both rosiglitazone and placebo groups [37]. Caution is needed in assessing CEP interpretation to ensure the validity of the evidence, especially if components of CEP vary in frequency, clinical meaningfulness and treatment effects [31,32,36,37,40]. Otherwise, one can always add on death in the composite, so long as it is sufficiently rare, to a lesser end point that shows statistical significance, and claim that the composite of mortality and the lesser end point have shown statistical significance [31,39,40].

Cordova *et al.* systemically reviewed composite end points of 40 RCTs. These authors summarize that components were often 'unreasonably combined, 'inconsistently defined' and inadequately reported. The result was confusion and exaggerated perception of the interventions [32]. The only factor about CEPs that is clear is that they are a double-edge sword; and researchers disagree about their interpretation. On one side CEPs can be advantageous to assess net effect of treatment and sometimes can help avoid bias, particularly in trials involving competing risks. For instance, Ferrera-Gonzalez *et al.* describe a primary cardiovascular study in which a (biased) conclusion of a significant treatment effect for fibrates versus a placebo on nonfatal AMI seems, at first, to be warranted. But also, there is a nonsignificant increase in a competing risk: death from any cause. Consequently, the real risk of AMI in those receiving fibrates is less because it is lowered in terms of patient-years due to the competing risk of mortality. However, if the CEP is formed of nonfatal AMI and mortality, the net effect is nonsignificant. This leads to a less-biased decision about the treatment effectiveness of the fibrates, possibly preventing interested stakeholders from being misled [31].

The disadvantage of CEP net effect can be seen in oncology trials, as the 'net effect' of the CEP makes it difficult to separate the disease-specific therapy effect that cancer trials seek in the presence of non-cancer competing risks [41]. The CEP used in oncology research is event-free survival end points where overall survival, disease-free survival and death from any cause, are often combined. In the presence of competing risks in oncology trials, misinterpretation, including misattribution of treatment effect can occur. For example, hormone therapy for prostate cancer reportedly interacts with age even without a cause-specific effect. If treatment affects competing risks such as non-cancer mortality and/or age, it might affect event-free survival, that is, the CEP net effect. The result of a CEP may simply be a consequence of covariates, other competing risks such as non-cancer mortality and other comorbidities, rather than a true treatment effect [41].

Multiple pathways exist in disease processes. Complicated oncology trial situations, especially with presence of many competing risks, call for adaptability and flexibility. Perhaps one resolution to this dilemma is a commonsense approach: examine the issue of treatment effectiveness from the perspective of a CEP, giving an overall net effect, and also from the perspective of separate multiple primary end points, facilitating a more detailed evaluation of the competing risks. This approach is rarely taken according to Mell and Jeong, perhaps owing to the difficulty of resolving conflicting conclusions from differing methods. Analyses should be conducted in both ways as it was in the fibrates study example provided above. In the end, the readers must keep in mind these conflicting interpretation modes of the CEPs [31,41].

From the design to interpretation of trial to avoid the 'unreasonably combined' CEPs, Berger argues, first to ensure that CEPs are "jointly fusible", forming an "information-preserving CEP" (IPCEP) [37]. An IPCEP should have a logical thread of connection among the components that allow them to cohere as one underlying composite variable. This simple

concept should stay with investigators at all times from design phase for careful construction of CEPs to data analyses and interpretation of the trial. The resulting IPCEPs are more likely to be homogenous in the frequency of events, effect sizes and clinical meaningfulness among the components. Hopefully any nonfusibility in an IPCEP will be an outlier rather than what is expected to be the norm, owing to careful construction of the components. Even if other symptom variables or health quality-of-life scales are added as secondary variables, the analysis of an IPCEP is still more straightforward and credible for the interpretation of treatment efficacy [42]. From a trial quality perspective then, it seems advisable for reviewers to set criteria to look for the joint fusibility of the components within, based on such description, definition and rationale for the selection of the IPCEPs in RCTs.

Examining the currently used tools for evaluating and reporting CEPs, Chalmers *et al.* and the CONSORT 2010 revision had questionnaire items concerning general end points, whether they were tabulated, and whether survival times and stopping dates were clearly provided in the study. Multiplicity issues of primary and secondary outcome were also mentioned. However, there was no mention of surrogate or CEPs. There were no criteria for examining rationale for logical selection of components in the composite, or what kind of relationship each component had with the clinical true end point [3,10,11]. Of note, in the Jadad scale, it completely fails to consider primary versus secondary or any type of end points, let alone the issues of multiplicity and heterogeneity described above [9]. Individual readers and reviewers of trials must understand that the highest ratings of the Jadad scale do not necessarily reflect high quality of a trial [1–3,42,43].

Spin reporting

Spin reporting is a reporting strategy that inaccurately highlights positive results and downplays negative findings. In the previous section we reviewed the influence of the heterogeneity within the composite potentially leading to misinterpretation of the result. The net benefit may be statistically significant, but the most important primary end point may not be. Knowingly and/or unknowingly, this misinformation can be reported to bolster one side of an argument. Even in large trials published in highly reputable journals, spin reporting remains common, as data are repackaged and marketed; that is, the negative outcome is censored and diluted by combining it in a composite [38,39].

There are various ways to spin a report. For example, not reporting the true clinical end point of mortality in a protocol is a common practice of reporting, showing 36% in Freemantle's systematic review [26]. One study compared conservative medical treatment with invasive revascularization surgery for older patients with chronic angina [37,44]. The CEP (major cardiac events) consisted of death, nonfatal MI, or hospital admission related ACS. The authors reported that revascularization greatly reduced the rate of major cardiac events compared with the medical treatment group, occurring in 72 (49%) of the medical group versus only 29 (19%) in the surgery group.

Further examination showed a marked difference in the frequency of hospital admissions; 75% of the admissions were in the medical treatment group, while death in the invasive treatment group was twice as high, which was not mentioned in the report [34]. By emphasizing an apparent statistical net benefit of the composite, the mortality report was essentially silenced [37,44]. Based on the positive spin of this trial report, patients who need to choose between conservative or surgical therapy may choose the surgery, without being aware of the increased risk of dying. They will have been malinformed and misled.

Even in masked trials, Ferreira-Gonzalez *et al.* note that the less important component of the composite events are emphasized instead of the primary end point, thus giving inaccurate perceptions of treatment efficacy to readers [40]. The net statistical effect of composite end point is a double-edged-sword phenomenon and needs to be treated as such [44–46].

In the most recent 2013 review, Vera-Badillo *et al.* examined the frequency of biased reporting of end points and toxicity in all published articles of Phase III breast cancer trials between 1995 and 2001 [47]. Of the 164 randomized clinical trials, 110 (67.1%) were biased in reporting of toxicity and 54 (32.9%) reported having positive treatment efficacy, despite a lack of statistical significance in the primary end point. Further analysis showed that when study results had higher p-values between the experimental and control groups, there was higher spin and biased reporting of toxicity, sometimes switching to secondary variables in order to imply benefits. By this spin, trial reports gave a false perception of efficacy and safety of the studied drug. This is no small matter from a clinical vantage point, considering how close the Phase III trials are to the public [45–48].

If we are to prevent and detect various types of spin, some additional features are clearly needed in currently available quality rating tools. With the current formats of the scale or checklist, it is farfetched to prevent or detect any sophisticated spin and bias in trial reports; it would be like trying to hit a bulls-eye whilst playing darts in the dark. Chalmers *et al.* and CONSORT would do well to include items for preventing

and detecting spin [10,11]. At a policy level, perhaps we should have a system that every protocol should be made public before data collection and analysis occur. A current trial report registry can contribute by mandating more specific criteria, to regulate transparency and detection of discrepancies before and after the data analyses. We believe Chalmers *et al.*'s rating scale and the CONSORT guide can improve their capacity to evaluate, guide, prevent and detect bias with the addition of trial quality features such as α preservation, SEPs, CEPs and spin report discussed here. The three criteria Jadad scale is too simplistic and limited for any further consideration regarding the issue [2,9].

Summarizing the answer to a question we had in mind for this article, 'can these quality rating tools indeed distinguish valid from flawed trials?', it seems that they fell short of an adequate evaluation or guide for trial quality. CONSORT [11] and Chalmers *et al.* [10] scales are more comprehensive and complete than the Jadad scale [9]. The Jadad scale leaves the critically unanswered questions of internal validity of trials it evaluates. The Jadad scale receives our respect for its low responder burden and the historical value, but capacity and function of a tool cannot be confused with historical value or convenience of its usage. Many reviewers caution against trials that received high or even perfect scores from Jadad ratings. The trial quality hardly reflects Jadad's claimed perfect scores [1–3,43,48] any more than we would rate a chess player by his or her ability to lift pieces off the board (physical strength), distinguish white squares from black squares (visual acuity), and reliably place the piece being moved on the desired square (manual dexterity). There is more to playing good chess and finding quality trials than just three elements. We believe it is time to let go of this archaic scale, recognizing that research methodology has evolved since the introduction of the Jadad scale in 1996.

Future perspective

Future quality rating instruments have to be developed from a completely different premise: the realization that a single flaw in any critical aspect of a trial can invalidate the entire trial [4]. This quality-rating tool will incorporate comprehensive criteria including, but certainly not limited to, the four features discussed in this paper. Furthermore, we propose that the ratings should incorporate a summary rating based on a multiplicative system offering scores between 0 and 100. This way, any trial that violates a criterion with serious flaws will receive a rating of 0, showing the internal validity of the trial to be invalid [2,49]. On the other hand, there are pragmatic aspects of flexible intervals of rating in less serious flaws in trial quality that are being developed and forthcoming. Quality rating criteria based on a multiplicative system will, we hope, offer tomorrow's researchers a fair pragmatic tool for guidance as they write their protocols for future trials [2]. Better quality rating criteria will also help meta-analysts, regulators, journal editors, funding agencies and other consumers of research to better distinguish valid trials from flawed ones, ultimately playing an important role in raising overall trial quality to a different dimension.

Financial & competing interests disclosure

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties. No writing assistance was utilized in the production of this manuscript.

Executive summary

- Four essential features among the many that affect trial quality were reviewed. These four features should be included when assessing randomized trials.
  - α preservation: commonly-used parametric analyses are misapplied in trials because data cannot be shown to be normally distributed. Instead, exact permutation tests should be employed; they are not difficult to conduct and can guarantee a preservation.
  - Surrogate end points: Despite their convenience and advantages, surrogate end points should not replace true clinical end points; even if they are validated, they should be used with extreme caution.
  - Composite end points: Although composite end points have some statistical advantages and popularity, they can lead to misinterpretation of trials. Close scrutiny of the components is mandatory.
  - Spin reporting: Surrogate and composite end points have been misused to highlight treatment benefits despite a lack of statistical significance, and to dilute negative findings of true clinical end points such as mortality or drug toxicity.
- Currently available tools for assessing trial quality fall well short of adequately protecting the public. The Jadad scale failed to address any of the features we discussed; we should not confuse any of its ratings with actual trial quality despite its historical value and (or, perhaps, because of) its low reviewer burden. The Chalmers *et al.* scale and the CONSORT checklist need to be updated with criteria for the issues discussed. For better search of evidence, quality rating instruments of the future should be comprehensive and updated to evaluate current trial methodology.

References

1 Berger VW. Selection bias and covariate imbalances in randomized clinical trials. John Wiley & Sons, Hoboken, NJ, USA (2005).

2 Berger VW, Alperson SY. A general framework for the evaluation of clinical trial quality. *Rev. Recent Clin. Trials.* 4(2), 79–88 (2009).

3 Palys K, Berger VW. A note on the Jadad Score as an efficient tool for measuring trial quality. *J. Gastrointest. Surg.* 17(6), 1170–1171 (2013).

4 Berger VW, Matthews JR. Conducting today's trials by tomorrow's standards. *Pharm. Stat.* 4, 155–159 (2005).

5 Berger VW. Pros and cons of permutation tests in clinical trials. *Statist. Med.* 19, 1319–1328 (2000).

6 Micceri T. The unicorn, the normal curve, and other improbable creatures. *Psychol. Bull.* 105(1), 156–166 (1989).

7 Geary R. Testing for normality. *Biometrika* 34, 209–242 (1947).

8 Berger VW. On the generation and ownership of α in medical studies. *Control. Clin. Trials* 25(6), 613–619 (2004).

9 Jadad AR, Moore RA, Carroll D *et al.* Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Control. Clin. Trials* 17, 1–12 (1996).

10 Chalmers TC, Smith HJ, Blackburn B *et al.* A method for assessing the quality of a randomized control trial. *Control. Clin. Trials* 2, 31–49 (1981).

11 Moher D, Hopewell S, Schulz K *et al.* CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *BMJ* 340, 332 (2010).

12 Institute of Medicine Committee on Qualification of Biomarkers and Surrogate End Points in Chronic Disease. *Washington, DC: Report on the Evaluation of Biomarkers and Surrogate End points in Chronic Disease.* The National Academies Press, Washington, DC, USA (2010).

13 Furgerson JL, Hannah WN, Thompson JC. Challenge of surrogate end points. *South Med. J.* 105(3), 156–160 (2012).

14 Bucher HC. Benefit and pitfalls in the use of data from surrogate end point trials for clinical decision making. *Evid. Fortbild. Qual. Gesundhwes.* 104(3), 230–238 (2010).

15 Friedman L. Data and Safety Monitoring Boards. In: Gallin J. and Ognibene, F. Principles and Practice of Clinical Research. Academic Press, London, UK (2007).

16 CAST Investigators. The Cardiac Arrhythmia Suppression Trial: Preliminary report: Effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. *N. Engl. J. Med.* 321(6), 406–412 (1989).

17 Packer M, Carver JR, Rodeheffer RJ *et al.* Effect of oral milrinone on mortality in severe chronic heart failure. *N. Engl. J. Med.* 325(21), 1468–1475 (1991).

18 Buyse MG, Molenberghs G, Burzykowski T *et al.* Statistical validation of surrogate end points: Problems and proposals. *Drug Inform. J.* 34(2), 447–454 (2000).

19 Buyse MG, Molenberghs G, Burzykowski T *et al.* The validation of surrogate end points in meta-analyses of randomized experiments. *Biostatistics* 1(1), 49–67 (2000).

20 Buyse M, Sargent DJ, Grother A *et al.* Biomarkers and surrogate end points – the challenge of statistical validation. *Clin. Oncol.* 7, 309–317 (2010).

21 Prentice RL. Surrogate end points in clinical trials: definition and operational criteria. *Stat. Med.* 8, 431– 440 (1989).

22 Begg CB, Leung DHY. On the use of surrogate end points in randomized trials. *J. Royal Stat. Soc.* 163, 15 –24 (2000).

23 Berger VW. Does the Prentice criterion validate surrogate end points? *Stat. Med.* 23(10), 1571–1578 (2004).

24 Buyse M, Molenberghs G. Criteria for the validation of surrogate end points in randomized experiments. *Biometrics* 54(3), 1014–1029 (1998).

25 Demets D, Califf R. Lessons learned from recent cardiovascular clinical trials: part I. *Circulation* 106, 746–751 (2002).

26 Freemantle N, Calvert M, Wood J, Eastaugh J, Griffin C. Composite outcomes in randomized trials: greater precision but with greater uncertainty? JAMA 289, 2554–2559 (2003).

27 Egger M, Smith GD, Altman DG. Systematic reviews on Healthcare: Meta-analysis in context 2nd ed. BMJ Publishing, London, UK (2007).

28 Naslund U, Grip L. Fischer-Hansen J *et al.* The impact of an end-point committee in a large multicentre, randomized, placebo-controlled clinical trial: results with and without the end-point committee's final decision on end-points. *Eur. Heart J.* 20, 771–777 (1999).

29 Palys K, Berger VW. On the incompleteness of CONSORT. *JNCI* 105(3), 244 (2013).

30 Tomlinson G, Detsky AS. Composite end points in randomized trials there is no free lunch. *JAMA.* 303(3), 267–268 (2010).

31 Buhr KA. Surrogate end points in secondary analyses of cardiovascular trials. *Prog. Cardiovasc.* 54(4), 343–350 (2012).

32 Ferreira-González I, Alonso-Coello P, Solà I *et al.* Composite end points in clinical trials. *Rev. Esp. Cardiol.* 61(3), 283–290 (2008).

33 Cordoba G, Schwartz L, Woloshin S *et al.* Definition, reporting, and interpretation of composite outcomes in clinical trials: systematic review. *BMJ* 341, c3920 (2010).

34 Gerstman BB. *Basic Biostatistics*. Joens and Barlett, Sudbury, MA, USA (2008).

35 Gordis L. *Epidemiology (4th edition)*. Saunders, Philadelphia, PA, USA (2004).

36 Hulley SB, Cummings SR, Browner WS *et al.* *Designing Clinical Research 3rd ed.* Lippincott Williams & Wilkins, Wolters Kluwer, Philadelphia PA (2007).

37 Montori VM, Permanyer-Miralda G, Ferreira-Gonzalez I *et al.* Validity of composite end points in clinical trials. *BMJ* 330, 594–596 (2005).

38 Kleist P. Composite end points for clinical trials: current perspectives. *Int. J. Pharm. Med.* 21, 187–198 (2007).

39 Wittkop L, Smith C, Fox Z *et al.* NEAT-WP4. Methodological issues in the use of composite end points in clinical trials: examples from the HIV field. *Clin. Trials* 7(1), 19–35 (2010).

40 Gerstein HC, Yusuf S, Bosch J, Pogue J *et al.* DREAM Trial: Effect of rosiglitazone on the frequency of diabetes in patients with impaired glucose tolerance or impaired fasting glucose: a randomised controlled trial. *Lancet* 368(9541), 1096–1105 (2006).

41 Ferreira-González I, Busse D, Heels-Ansdell V *et al.* problems with use of composite end points in cardiovascular trials: systematic review of randomised controlled trials. *BMJ* 334(7597), 786–788 (2007).

42 Mell L, Jeong J. Pitfalls of using composite primary end points in the presence of competing risks. *J. Clin. Oncol.* 28(28), 4297–4298 (2010).

43 Berger VW. Improving the information content of categorical clinical trial end points. *Control. Clin. Trials* 23, 502–514 (2002).

44 Alperson SY, Berger VW. Opposing systematic reviews: the effects of two quality rating instruments on evidence regarding T'ai Chi and bone mineral density in post-menopausal women. *J. Altern. Complement. Med.* 17(5), 389–395 (2011).

45 TIME Investigators. Trial of invasive versus medical therapy in elderly patients with chronic symptomatic coronary-artery disease: a randomised trial. *Lancet* 358(9286), 951–957 (2001).

46 Boutron I, Douron S, Ravaud P *et al.* Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *J. Am. Med. Assoc.* 303, 2058–2064 (2010).

47 Ocana A, Tannock IF. When are 'positive' clinical trials in oncology truly positive? *J. Natl Cancer Inst.* 103, 16–20 (2010).

48 Vera-Badillo FE, Shapiro R, Ocana A *et al.* Bias in reporting of end points of efficacy and toxicity in randomized, clinical trials for women with breast cancer. *Ann. Oncol.* 24(5), 1238–1244 (2013).

49 Palys KE, Berger VW, Alperson S. Trial quality checklists: on the need to multiply (not add) scores. *Clin. Oral. Investig.* 17(7), 1789–1790 (2013).