American Journal of Clinical and Experimental Medicine
2013; 1(1): 24-34
Published online July 20, 2013 (http://www.sciencepublishinggroup.com/j/ajcem)
doi: 10.11648/j.ajcem.20130101.15

Composite endpoints: Sometimes more than a solely economic consideration

Charles J Kowalski

University of Michigan, Ann Arbor MI 48109 USA

Email address:
chuckk@umich.edu

To cite this article:
Charles J Kowalski. Composite Endpoints: Sometimes More than a Solely Economic Consideration. *American Journal of Clinical and Experimental Medicine.* Vol. 1, No. 1, 2013, pp. 24-34. doi: 10.11648/j.ajcem.20130101.15

Abstract: Endpoints are response variables, or outcomes, that are measured during the course of a clinical trial. I consider endpoints that are either events (e.g., death) or the time to an occurrence of an event (e.g., time to disease progression). A composite endpoint (CEP) is an endpoint that consists of a number of component endpoints, and is considered to have occurred as soon as any one of its components occurs. For example if CEP = death + disease progression, the CEP is said to have occurred as soon as either the disease progresses or the patient dies. It is seen that one of the results of using a CEP is to increase the event rate; and this in turn can reduce the sample size or the time required to observe a specified number of events, thereby resulting in a speedier, less costly clinical trial. Many believe that the *only* reason CEPs are ever employed is to this end, viz., saving money. I argue that there may be other circumstances that suggest the use of CEPs – that the choice of the primary response variable should be driven by the question the trial is being designed to answer.

Keywords: Clinical Trials, Outcome Variables, Event Rates, Win Ratio, Quality of Life

1. Introduction

Many clinical trials are designed to compare two or more interventions by following subjects for a period of time and comparing event rates in the groups. For ease in exposition, I consider but two interventions, A and B, and often use the language of drugs instead of the more general 'interventions,' but conclusions reached in this context will often admit more general applicability. The *events* whose rates are to be compared may be quite varied. However, if we restrict attention to phase III clinical trials designed to determine *what drug should be used in clinical practice* (Kowalski 2010), mortality and events related to quality of life (QoL) are obvious candidates (Kowalski et al 2008, 2012). Often, mortality will be recognized as the most relevant primary outcome, but in many situations large samples and long periods of time will be required for this outcome to be realized in sufficient numbers to be amenable for analysis, thereby increasing the cost of the study. In an attempt to increase the baseline event rate for the primary outcome variable (so as to reduce the required sample size and/or the time to observe a specified number of events), some have suggested the use of composite endpoints (CEPs). CEPs are defined as the occurrence of any event among a given set of events after a certain period of follow-up (Ferreira-Gonzalez et al 2007) and these are often used in phase III clinical trials (Freemantle et al 2003).[1] For example, Braunwald et al (1992) used the "Unsatisfactory Outcome" endpoint that involved ten components, including mortality, several nonfatal endpoints (e.g., recurrent infarction, congestive heart failure, left ventricular dysfunction), and several safety endpoints (e.g., intracranial hemorrhage, anaphylaxis). The CEP, used as the primary endpoint in the trial, was the presence of *any one* of these events. Here, as is often the case, all-cause mortality is a component of the CEP but the inclusion of additional variables makes interpretation more difficult (Freemantle et al 2003). In particular, significance of the CEP does not demonstrate efficacy in the individual components of the CEP.

In this paper, I discuss a number of the issues surrounding the use of CEPs in phase III clinical trials. While I do not favor the use of CEPs when advanced *solely* to save time and

[1] Alternatively, one could use time to first occurrence of any event in the set. This would result in a continuous, rather than a discrete outcome. For definiteness, I focus on the simple occurrence of one of the events and the use of relative risk to compare the groups. On occasion, however, I will consider 'time to first event' where survival analysis statistical techniques (e.g., hazard ratios) are used to compare the groups.

money, [2] this does not mean there cannot be more substantive reasons for their use in certain situations. As I have maintained before (Kowalski 2010; Kowalski and Mrdjenovich 2013) *the design of a clinical trial should be driven by the clinical question being addressed* – and the choice of outcome is a critical aspect of trial design. I do not intend to belabor the details of every step of the argument in favor of this principle, but I will start at the beginning and sketch some of the more important points. I follow what is often called the traditional (as opposed to Bayesian) approach to clinical trial design. See, e.g., Friedman, Furberg and DeMets (1998).

Many clinical trials are designed to compare two or more interventions by following subjects for a period of time and comparing event rates in the groups. For ease in exposition, I consider but two interventions, A and B, and often use the language of drugs instead of the more general 'interventions,' but conclusions reached in this context will often admit more general applicability. The *events* whose rates are to be compared may be quite varied. However, if we restrict attention to phase III clinical trials designed to determine *what drug should be used in clinical practice* (Kowalski 2010), mortality and events related to quality of life (QoL) are obvious candidates (Kowalski et al 2008, 2012). Often, mortality will be recognized as the most relevant primary outcome, but in many situations large samples and long periods of time will be required for this outcome to be realized in sufficient numbers to be amenable for analysis, thereby increasing the cost of the study. In an attempt to increase the baseline event rate for the primary outcome variable (so as to reduce the required sample size and/or the time to observe a specified number of events), some have suggested the use of composite endpoints (CEPs). CEPs are defined as the occurrence of any event among a given set of events after a certain period of follow-up (Ferreira-Gonzalez et al 2007) and these are often used in phase III clinical trials (Freemantle et al 2003). [3] For example, Braunwald et al (1992) used the "Unsatisfactory Outcome" endpoint that involved ten components, including mortality, several nonfatal endpoints (e.g., recurrent infarction, congestive heart failure, left ventricular dysfunction), and several safety endpoints (e.g., intracranial hemorrhage, anaphylaxis). The CEP, used as the primary endpoint in the trial, was the presence of *any one* of these events. Here, as is often the case, all-cause mortality is a component of the CEP but the inclusion of additional variables makes interpretation more difficult (Freemantle et al 2003). In particular, significance of the CEP does not demonstrate efficacy in the individual components of the CEP.

In this paper, I discuss a number of the issues surrounding the use of CEPs in phase III clinical trials. While I do not favor the use of CEPs when advanced *solely* to save time and money, [4] this does not mean there cannot be more substantive reasons for their use in certain situations. As I have maintained before (Kowalski 2010; Kowalski and Mrdjenovich 2013) *the design of a clinical trial should be driven by the clinical question being addressed* – and the choice of outcome is a critical aspect of trial design. I do not intend to belabor the details of every step of the argument in favor of this principle, but I will start at the beginning and sketch some of the more important points. I follow what is often called the traditional (as opposed to Bayesian) approach to clinical trial design. See, e.g., Friedman, Furberg and DeMets (1998).

2. What Is the Question

The design of a clinical trial depends on the question that the clinical investigator is addressing. It is recognized that typically more than one – even many – questions will be of interest to the investigator, but it is necessary to choose one of these as *primary* – *the* question that the trial will be designed to answer. Stating this question clearly, and in advance, is necessary for the development of the design of the study (Friedman et al 1998, 16). For one thing, the required sample size for the study is what is necessary to answer the primary question. Response variables are outcomes (endpoints) measured during the course of the trial, and their choice is intimately related to the study question: these outcomes help define and answer the question. In a simple treatment/control comparison for example, the sample size required will involve, among other things, the difference between the chosen primary response variable, and the magnitude of this difference considered to be of clinical moment. O'Brien and Geller (1997, 222) pointed to the importance of aligning trial outcomes and their analysis to the medical questions that the trial is designed to answer. They showed how "perfectly reasonable test procedures can lead to absurd results when they are matched to the wrong medical questions." One *first* clearly states the clinical question, *then* chooses outcomes that match, and *then* determines an appropriate analysis strategy. Should the outcome of *this* process contain ambiguities, their root cause will usually be discovered by questioning the question. Job 1 for an outcome measure is that it be chosen so as to enable answering the question posed. It will be helpful if it also has some additional properties, considered next.

Ideally, the primary end point should be clinically relevant,

[2] Zivin (2000) noted that clinical trials can be described in three ways: *trustworthy*, *fast*, or *cheap*; and a given trial can have only two of these characteristics. Aiming for *fast* and *cheap* seems misguided, at best.

[3] Alternatively, one could use time to first occurrence of any event in the set. This would result in a continuous, rather than a discrete outcome. For definiteness, I focus on the simple occurrence of one of the events and the use of relative risk to compare the groups. On occasion, however, I will consider 'time to first event' where survival analysis statistical techniques (e.g., hazard ratios) are used to compare the groups.

[4] Zivin (2000) noted that clinical trials can be described in three ways: *trustworthy*, *fast*, or *cheap*; and a given trial can have only two of these characteristics. Aiming for *fast* and *cheap* seems misguided, at best.

be easily ascertainable in all patients, be capable of unbiased assessment, be sensitive to the hypothesized effects of the treatment, and be inexpensive to measure (Neaton et al 2005, 568). Different stakeholders [5] may weight these characteristics differently. For example, this last characteristic ($s) is the least important scientifically, but the same cannot be said for trial sponsors. Sponsors will want to save money. One approach to saving money is to use a *surrogate outcome*, one that stands-in for the primary outcome, but is less costly to obtain. In a recent paper (Kowalski 2013), I discussed the use of surrogate outcome measures in phase III clinical trials of new drug products and suggested that their use was seldom – if ever – justified. The basic idea is that the *only* reasons for taking drugs are either to prolong life or increase its quality (or both); thus the only directly relevant outcomes are mortality and/or quality of life. The practical problem with this view is that trials with mortality as outcome will often require a substantial time (and hence monetary) commitment, and so trial sponsors have incentives for choosing outcomes that can be assessed more readily. CEPs are not necessarily surrogates, even though one of the major reasons given for their use is to cut expenses. In addition, other reasons have been advanced to justify the use of CEPs. Thus their possible use requires consideration over and above that given to surrogates. Using cholesterol level *instead* of mortality is using a surrogate outcome. Many CEPs, on the other hand, will *include* mortality and so are not surrogates in the usual sense (of *replacement by*).

The use of CEPs is contentious. Some have focused on their purported advantages; others have pointed to their limitations. Ferreira-Gonzàles et al (2007) summarized this material, listing some of the claimed advantages and disadvantages of CEPs. Among the advantages attributed to CEPs are:

- Reduces sample size requirement (A1);
- Estimates the net clinical benefit of a therapy (A2);
- Improves understanding of the effect of the interventions avoiding competing risks (A3);
- Avoids the need to choose a single primary endpoint when many may be of equal importance (A4); and
- Avoids adjustment for multiple comparisons (A5).
- Disadvantages attributed to CEPs include:
- Practical interpretation could be problematic when component endpoints are dissimilar in patient importance (D1);
- Interpretation can be problematic if either the event rates or relative risk reduction vary appreciably across components (D2);
- Potential masking of an increase in a harmful effect associated with an experimental intervention (D3);
- Possibility of biases secondary to competing risk (D4);
- The larger the number of components the more work to accurately ascertain the composite (D5);
- Excessive influence of the more subjective (clinician-driven) component outcomes (D6); and
- Alpha error must be adjusted to draw confirmatory conclusions about the components (D7).

The trick is to use CEPs when and only when there is a favorable advantage/disadvantage ratio. Any chance of actually being able to do this in practice will necessarily involve the detailed particulars of the trial under consideration and its motivating question. It will also involve a closer look at the advantages and disadvantages listed above. It will not be necessary to consider each item in both lists separately, e.g., A3 and D4 both address competing risks (yes, some think CEPs advantageousThe design of a clinical trial depends on the question that the clinical investigator is addressing. It is recognized that typically more than one – even many – questions will be of interest to the investigator, but it is necessary to choose one of these as *primary – the* question that the trial will be designed to answer. Stating this question clearly, and in advance, is necessary for the development of the design of the study (Friedman et al 1998, 16). For one thing, the required sample size for the study is what is necessary to answer the primary question. Response variables are outcomes (endpoints) measured during the course of the trial, and their choice is intimately related to the study question: these outcomes help define and answer the question. In a simple treatment/control comparison for example, the sample size required will involve, among other things, the difference between the chosen primary response variable, and the magnitude of this difference considered to be of clinical moment. O'Brien and Geller (1997, 222) pointed to the importance of aligning trial outcomes and their analysis to the medical questions that the trial is designed to answer. They showed how "perfectly reasonable test procedures can lead to absurd results when they are matched to the wrong medical questions." One *first* clearly states the clinical question, *then* chooses outcomes that match, and *then* determines an appropriate analysis strategy. Should the outcome of *this* process contain ambiguities, their root cause will usually be discovered by questioning the question. Job 1 for an outcome measure is that it be chosen so as to enable answering the question posed. It will be helpful if it also has some additional properties, considered next.

Ideally, the primary end point should be clinically relevant, be easily ascertainable in all patients, be capable of unbiased assessment, be sensitive to the hypothesized effects of the treatment, and be inexpensive to measure (Neaton et al 2005, 568). Different stakeholders [6] may weight these characteristics differently. For example, this last characteristic ($s) is the least important scientifically, but the same cannot be said for trial sponsors. Sponsors will want to save money. One approach to saving money is to use a

[5] Stakeholders may include sponsors, subjects, investigators, IRBs, DSMBs, regulatory authorities (FDA), medical providers and patients.

[6] Stakeholders may include sponsors, subjects, investigators, IRBs, DSMBs, regulatory authorities (FDA), medical providers and patients.

surrogate outcome, one that stands-in for the primary outcome, but is less costly to obtain. In a recent paper (Kowalski 2013), I discussed the use of surrogate outcome measures in phase III clinical trials of new drug products and suggested that their use was seldom – if ever – justified. The basic idea is that the *only* reasons for taking drugs are either to prolong life or increase its quality (or both); thus the only directly relevant outcomes are mortality and/or quality of life. The practical problem with this view is that trials with mortality as outcome will often require a substantial time (and hence monetary) commitment, and so trial sponsors have incentives for choosing outcomes that can be assessed more readily. CEPs are not necessarily surrogates, even though one of the major reasons given for their use is to cut expenses. In addition, other reasons have been advanced to justify the use of CEPs. Thus their possible use requires consideration over and above that given to surrogates. Using cholesterol level *instead* of mortality is using a surrogate outcome. Many CEPs, on the other hand, will *include* mortality and so are not surrogates in the usual sense (of *replacement by*).

The use of CEPs is contentious. Some have focused on their purported advantages; others have pointed to their limitations. Ferreira-Gonzàles et al (2007) summarized this material, listing some of the claimed advantages and disadvantages of CEPs. Among the advantages attributed to CEPs are:

- Reduces sample size requirement (A1);
- Estimates the net clinical benefit of a therapy (A2);
- Improves understanding of the effect of the interventions avoiding competing risks (A3);
- Avoids the need to choose a single primary endpoint when many may be of equal importance (A4); and
- Avoids adjustment for multiple comparisons (A5).

Disadvantages attributed to CEPs include:

- Practical interpretation could be problematic when component endpoints are dissimilar in patient importance (D1);
- Interpretation can be problematic if either the event rates or relative risk reduction vary appreciably across components (D2);
- Potential masking of an increase in a harmful effect associated with an experimental intervention (D3);
- Possibility of biases secondary to competing risk (D4);
- The larger the number of components the more work to accurately ascertain the composite (D5);
- Excessive influence of the more subjective (clinician-driven) component outcomes (D6); and
- Alpha error must be adjusted to draw confirmatory conclusions about the components (D7).

The trick is to use CEPs when and only when there is a favorable advantage/disadvantage ratio. Any chance of actually being able to do this in practice will necessarily involve the detailed particulars of the trial under consideration and its motivating question. It will also involve a closer look at the advantages and disadvantages listed above. It will not be necessary to consider each item in both lists separately, e.g., A3 and D4 both address competing risks (yes, some think CEPs advantageous, others not so).

3. CEP Advantages

Consider first the sample size/power advantage. Since the-more-events-the-better is generally true in this context, A1 is rarely challenged, but even here there can be extenuating circumstances. Neaton et al (2005, 568) put it this way: "The primary rationale for considering a composite primary outcome instead of a single event outcome is sample size. In success/failure trials and time to event trials, a higher event rate can lead to a smaller sample size or trial duration. 'Can' is an important choice of words. In some cases, power can be lost if the treatment does not affect, or affects to a lesser degree, 1 or more components of the composite end point." If the relative risk reduction (RRR) for a proposed new component is similar to the RRR for the event already being considered, then power/sample size requirements *will* go in an advantageous direction. For example, following Neaton et al (2005), consider an outcome for which the control event rate is 10% and the experimental event rate is 5% (RRR = (10 -5)/10 = 50%). If we fix the level of significance at 5% (two-sided) and the power at 90%, the required sample size is 1170 (585 per group). If one adds a component with the same RRR, say 20% and 10% control and experimental rates, so that the resulting composite has rates 30% and 15%, the sample size needed is reduced to 330 (115). If, on the other hand, we consider adding a new component to the original outcome that results in a smaller RRR, viz., 20% and 17.5%, so that for the new composite, RRR = (30 – 22.5/30 = 25%, the sample size *increases* to 1450 (725). The problem here is that, even though the total number of events is increased by using the composite, the difference between groups has been decreased, making them more difficult to distinguish. Another example is given by Ferreira-Gonzàlez et al (2008). The bottom line here is that while the use of a CEP may reduce the sample size requirement, it is not *guaranteed* to do so; what is guaranteed is that single-minded focus on *faster* and *cheaper* will impinge on the trustworthiness of the trial (Zivin 2000).

Getting at the "net effect of an intervention" may be important when significant risks accompany the beneficial effects of an intervention. Ferreira-Gonzàlez et al (2008) give the example where a new thrombolytic agent is under investigation for treatment of acute myocardial infarction (AMI). It is known that thrombolytic therapy increases the risk of cerebral hemorrhage, and it is thought that the new thrombolytic agent will reduce mortality, but perhaps at the expense of a slightly higher risk of hemorrhage as compared to standard thrombolytic therapy. Since the primary aim of thrombolytic therapy is to reduce mortality without increasing untoward side effects, one way to compare the treatments would be to use the CEP = mortality + cerebral hemorrhage. In order to outperform standard therapy, the expected decrease in mortality could not be offset by an

increase in hemorrhage. The two components of the composite may point in different directions, and the total number of events will temper a decrease in one by an increase in the other. The "net effect" in this example is the reduction in mortality corrected by any increase in cerebral hemorrhage. Ferreira-Gonzàlez et al (2008) suggest that, "A simple strategy for assessing the effect of interventions associated with important clinical risks is using a CEP that combines 'efficacy' and 'safety' outcomes. If the new intervention leads to a statistically significant decrease in the percentage of the events making up the CEP, we can be certain that this intervention is, in general, more beneficial than the standard one" (p. 285). There is an important proviso to this last statement which was in fact recognized by Ferreira-Gonzàlez et al (2008), but did not cause them to alter the quoted text. The 'generally more beneficial guarantee' can be made only if the components of the CEP are, at least roughly, equally important. They consider the CEP = death + cerebral hemorrhage + new pathology Q waves in the electrocardiogram. Imagine now that the new treatment has but little effect on mortality, causes an increase in cerebral hemorrhages, but has a marked beneficial effect on new pathological Q waves. The inclusion of the Q waves tilts the purported advantage in the direction of the new treatment, but it does so only because Q waves have been put on an equal footing with hemorrhages. One way to handle this problem is to not let it happen – insist that the components of the CEP *be* similar with respect to clinical impact. This will not always allow us to answer the clinical question as posed, and some way to incorporate components with different impacts will need to be employed. I return to this question later.

A competing risk is an event that removes a subject from being at risk for an outcome under investigation. Consider the event 'nonfatal AMI' which might be of interest if there is reason to believe a new intervention can be beneficial in its prevention, and has in fact been used as an endpoint in several cardiology clinical trials (Fleiss et al 1990). It is clear that should a subject die (from any cause) s/he is no longer able to suffer a nonfatal AMI; death is a competing event. If this is not accounted for in some way, a trial comparing the new with the old may (say by chance alone, or because there is additional toxicity associated with the new treatment) experience more deaths in the treatment group than in the control. This could bias the results in favor of the new treatment inasmuch as fewer subjects in that group would be at risk for nonfatal AMI. But even more ambiguity exists. Fleiss et al (1990) pointed out that the use of nonfatal AMI alone could lead to the following situation: A *reduction in nonfatal AMI could be harmful* if the treatment has no effect on the overall incidence of AMI but increases their severity so that more are fatal. On the other hand, a *reduction in nonfatal AMI could be beneficial* if the intervention reduces the overall incidence of AMIs without influencing those that are fatal. Fleiss et al (1990, 684) concluded that "when there is research interest in nonfatal infarctions, they should be analyzed in tandem with fatal infarctions and not by themselves." One way to do this is to use the CEP = nonfatal infarction + fatal infarction, but other in-tandem analysis strategies are possible. Ferreira-Gonzàlez et al (2008) suggested the use of CEP = death + nonfatal AMI. Notice that while Ferreira-Gonzàlez et al (2008) used all-cause mortality in their CEP, Fleiss et al (1990) focused on fatal infarcts. I will have more to say about the distinction between all-cause and disease-specific mortality later; here I emphasize only that Fleiss et al (1990) said *when there is a research interest in nonfatal infarctions* reinforcing my earlier point that trial design (including endpoint selection) should be driven by the question we expect the trial to address.

I will consider A4 and A5 together as they are both addressed by the ICH (1999) Guidance on Clinical Trials, which states "If a single primary variable cannot be selected from multiple measurements associated with the primary objective, another useful strategy is to integrate or combine the multiple measurements into a single or 'composite' variable using a predefined algorithm … This approach addresses the multiplicity problem without requiring adjustment to the type 1 error" (p. 1911-12).

The first of these – choosing one from the many – is apt to be more problematic in pragmatic (as opposed to explanatory) clinical trials (Kowalski 2010). After noting that "the choice of a small number of criteria [outcomes] is the mark of a clearly formulated hypothesis," one of the defining characteristics of the explanatory approach, Schwartz et al (1980, 49) pointed out: "With a pragmatic approach the situation is quite different. Here we must take account of all the practically important criteria and there may well be many of these. However, at the analysis stage they cannot be considered singly, for only one decision can be taken and this must rest on an overall balance of the advantages and disadvantages." Combining many measurements into one is, in general, a difficult problem, beset with difficulties in interpretation. I pointed to some of these some 40 years ago (Kowalski 1972); surely some progress has since been made, see, e.g., Zhang et al (1997), but predefining that algorithm can still present formidable challenges. Use of a CEP might work if "a single primary variable cannot be selected from multiple measurements associated with the primary objective" because they are really considered to be 'equally important,' but this seems unlikely if even just one stakeholder's view is adopted, all but impossible if multiple stakeholders are to be satisfied. It needs to be remembered that sponsors will design trials that will efficiently and economically allow them to market a drug. Subjects would like to avoid invasive outcome measures and/or drugs with serious side effects. Investigators will prefer techniques and measures with which they are familiar. IRBs will be juggling risks and benefits. The FDA must be focused on safety and efficacy. Physicians will want something their patients will take and that will make them feel better. Patients want both a better QoL and more of it. It's a wonder that these factions can agree on *anything*, let alone on one, single measure to serve

as the primary outcome in a clinical trial.

With regard to the multiplicity problem, a CEP is a solution only if significance (or not) of the CEP is *enough* to satisfy the trial's aim. Soon as one enquires into the significance of individual components of the composite, the multiplicity problem must be faced. FDA (2006, 28) Guidance notes "Though one reason for the use of a composite is to reduce the multiplicity problems associated with multiple separate endpoints, composites can do so only if it is agreed that treatment impact on each of the endpoints is of value and if the endpoints move in the same direction" … "Multiplicity problems arise when the multiple individual components of a composite endpoint are intended as possible claims. In general, individual components of a composite measure will not be adequate to support a claim unless the components are prespecified in the SAP [Statistical Analysis Plan] as separate endpoints, either sharing overall study alpha (co-primary endpoints) or identified in a sequential analysis, and the study results are found statistically and clinically meaningful in the context of the total composite and other individual component results." Sponsors will be anxious to market their products for as many indications as possible, but product labeling will point only to those of demonstrated efficacy.

4. CEP Disadvantages

Interpretation (D1 and D2) of CEPs can be problematic. These problems can be mitigated by simply taking to heart the stated cautions in D1 and D2: *When possible*, choose components that are similar with respect to import-to-patients and can be expected to have similar event rates and/or relative risk reductions. Montori et al (2005) suggested asking the three questions: (1) Are the component end points of similar importance to patients? (2) Did the more and less important end points occur with similar frequency? (3) Are the component end points likely to have similar relative risk reductions? Affirmative answers to all three questions would presumably allow one to focus on the CEP, without being required to examine the components separately. There are several reasons to tread carefully here. For one, a 'yes' to (2) would mean more *and* less important end points exist, so that the answer to (1) would have to be 'no.' Reflection shows that, in many situations, it will not be possible to fill the CEP stocking with component goodies that satisfy these criteria. If we are to design trials (including endpoint selection) to answer realistic clinical questions, it may be necessary to include a number of outcomes, some of which will necessarily be of more importance to some patients than others (or of more importance to a given patient at another point in time), and will not all respond with the same sensitivity to treatment. Clinical reality is usually complex. Chi (2005, 609) noted, "There are often many clinically relevant and important endpoints or variables that are needed to fully characterize a disease and to properly assess the effect of a treatment on the disease … a disease may be characterized by its pathophysiology, severity, signs and symptoms, progression, morbidity, mortality, etc." Montori et al (2005, 596) thought that CEPs would not be of much use in such situations, stating "The validity of composite end points depends on similarity in patient importance, treatment effect, and number of events across the components … When large variations exist between components the composite end point should be abandoned." I believe that, while there are times when the use of a CEP will be contraindicated, their use need not be abandoned unless the components are all-but-interchangeable. Indeed, there are several approaches that may prove fruitful in such situations. These are described below in the section on Weighted/Hierarchical Approaches. I continue here with the remaining disadvantages attributed to CEP usage.

Coping with D3 and D4 will often involve including as components in the CEP events that are clearly different in importance to patients, but properly recognized as such, and appropriately accounted for in the analysis. We can mask an important adverse event (hemorrhage) if we include Q waves, but this will not occur if we account in some way for the fact that these events are not clinically interchangeable. This is possible and discussed in the Weighted/Hierarchical section. There too will be found strategies to handle the mismatch in importance between death and a non-fatal AMI that necessarily accompanies the inclusion of death in the CEP to avoid a bias due to a competing risk. D1 – D4, then, are all dealt with, when necessary, by not treating unequals as equal, but by appropriately accounting for any inequalities in the design and analysis of the trial.

D5 says only that the more components in a CEP, the more work it is to accurately assess the CEP. It is hard to gainsay what many would consider a near-tautology, but it should come as no particular surprise that it may be more difficult to answer a hard question than an easy one. The dimension of a CEP – and even if a CEP should be used at all – is best determined by the clinical question. Useful answers to important questions will not often be obtained without significant effort.

D6 speaks to "excessive influence of subjective components." Use of the word *excessive* says it all – of course, we would not want the influence of *any* component to be *excessive*. We are again directed to pursue strategies that will allow individual components to exert *appropriate* influence on interpretation. It will not do to ignore the information, albeit subjective, available from, e.g., QoL assessments as these are usually of great importance to patients. How to incorporate such information is discussed in the following section. I will note here only the danger(s) inherent in allowing conflicts of interest to influence subjective judgments that may dominate overall CEP assessment.

D7 alerts us to the fact that many analysts consider it necessary to adjust levels of significance should we want to test hypotheses involving individual components of a CEP. I begin by noting that not all schools of statistical thought are convinced of any need to do any adjustment at all (Rothman, 1990; Berry and Hochberg, 1999), suggesting instead that,

"statistical adjustments for multiplicity provide crude answers to irrelevant questions" (Schultz and Grimes, 2005, 1591). This is a controversy that cannot be considered with due diligence here, but it should be noted that even if one is convinced of the need to do adjustments/corrections, this hardly counts as a *disadvantage*. It is the price that must be paid to ensure that the components are assessed with the care needed as dictated by traditional views of statistical inference. I also think it is important to recognize that different stakeholders may well have differing attitudes concerning the approach taken to assess the individual components of a significant CEP. As put by Freemantle et al (2003, 2558), "It is particularly in the dissection of a composite that the differing interests of sponsors, licensing authorities, and interpreters become manifest."

5. Weighted/Hierarchical Approaches

If the clinical question being addressed requires the use of components differing in importance, it is possible to assign weights to the components reflecting relative importance. Neaton et al (2005) give several examples of this, and point to the importance of validating the weights by relating the weighted composite to a credible global outcome. The new treatment response can then be compared to the control using the Mann-Whitney rank sum test. One advantage of this approach is the intuitively satisfying interpretation attached to the associated test statistic. The probability that a randomly selected patient from the new treatment group will respond better than a randomly selected patient receiving standard therapy is estimated by U/mn, where U is the Mann-Whitney test statistic and m, n are the respective sample sizes (O'Brien and Geller 1997). The most difficult aspect of this approach is getting at which weights to use. Subjective assignment is always possible, but is open to obvious criticisms, and will require further analysis to see how much inferences depend on the weights chosen (sensitivity analysis). Cardiff et al (1990) had a better idea; they asked a number of cardiologists to rank outcomes from 0 – 10 on endpoints that might be affected by a reperfusion strategy. However obtained, before use, weights might be validated by using them in completed trials in which the various endpoints were assessed, but not weighted in the primary analysis. Neaton et al (1994) did just that in a number of completed HIV treatment trials. It should be noted that while weighting presents nontrivial challenges, the alternative of considering all components to be weighed equally, will present its own set of challenges, which promise to be even more serious than moderate weight misspecification.

Lubsen and Kirwan (2002) did not assign weights, but noted that different *kinds* of trial outcomes could be ranked hierarchically, viz.,

- Level 1 – All-cause mortality
- Level 2 – Cause-specific mortality
- Level 3 – Non-fatal clinical events
- Level 4 – Symptoms, signs and paraclinical measures

Problems can occur when the analysis of an endpoint other than all-cause mortality ignores information from higher levels, e.g., if non-fatal AMIs are analyzed without recognizing the fact that these can occur only in the living, or if QoL is assessed without taking into account previous non-fatal AMIs. Examples were given showing how composite endpoints with components selected to appropriately represent the different levels aided interpretation. Of particular interest is the simple example they give involving hospitalization (Level 3) as the event of interest. Recognizing that death (Level 1) is a competing event, they consider the CEP = death + hospitalization, and they show that it is important to collect and display the data so that all possible outcomes are considered and that they are displayed in mutually exclusive categories as below, where (X, Y) represents the observed %s in the active treatment and control groups:

- D and H (15, 5)
- D and not H (5, 15)
- A and H (20, 20)
- A and not H (60, 60)

If one were to look at only the total number of deaths (20, 20) and hospitalizations (25, 35), it would be tempting to conclude that while active treatment did not affect mortality, it did have a favorable effect on hospitalization. To stop here would, however, ignore the clinically relevant information contained in A and not H, *hospitalization-free survival*, which is the same in both groups. When this is factored in to the interpretation, it appears that the hospitalization advantage exists only because more died while in the hospital and not before admission. It is recognized that this example, using contrived numbers to illustrate a point, may be considered unrealistic by some. Lubsen and Kirwan (2002) show, however, that this problem was also realized in the dofetilide trial, where a drug to stabilize heart rhythm was compared to placebo in some 1500 patients with congestive heart failure (Torp-Pedersen et al 1999). These examples clearly illustrate the importance of collecting the data needed to consider all possible combinations of outcomes in the CEP, and the usefulness of measures like hospitalization-free survival in interpreting the results. Data collection is a design consideration that needs to be determined prior to initiating the trial. Chi (2005, 612) noted that, "Failure to collect information on potential fatal and other non-fatal events, after a patient has reported a first non-fatal event, renders the analysis of individual component endpoint difficult to interpret. The reason is that such individual component analysis is based on incomplete and censored data where the censoring is likely to be treatment dependent [informatively censored]." … "It is recommended that the study design should require patients who had a first non-fatal event to remain in the trial under treatment, if possible, and continue follow-up and measurement of all component outcomes till the end of the study." If patients remain in the trial after reporting a non-fatal event but are crossed-over to another treatment, the resulting analysis will be challenging, but in many situations

this problem will have to be faced because it will be unethical to continue a patient on a treatment that has failed. One way to circumvent the crossover problem is to use an outcome like time-to-treatment-failure (TTF) which is recorded before the crossover is even made.

Pocock et al (2012) recently developed a method for dealing with composite endpoints that is easy to use and gives appropriate priority to the more clinically relevant components. It can be used when the time-to-event form of CEP is envisaged. They consider a simple example with components cardiovascular death (CVD) and hospitalization for chronic heart failure (HHF). Recognizing that CVD is more important than HHF, patients matched on the basis of their risk profiles, who were randomly assigned to either the new treatment or control, are compared first with respect to time to CVD, determining whether either one had a CVD before the other. If one did, the longer lasting member of the pair is declared the 'winner.' If not, only then does one proceed to compare HHF times. The one with the longer time to HHF is the 'winner.' Then, concentrating on the new treatment group, we count the number of winners and losers (ignoring ties for the moment), and the *win ratio* is #W/#L. Inference is based on the proportion #W/(#W + #L) for which confidence intervals and tests of significance can be derived. The proportion of pairs that were tied (e.g., if neither suffered either event) is considered to be a useful supplementary statistic. I consider the matched-pairs nature of this approach to be one of its strong points, but patients need not be matched to employ the technique. The computations are more extensive (e.g., each patient in the treatment group is compared to each patient in the control), but are still based on the simple counting of winners and losers. The approach can extend to more than two components as long as they can be sensibly ordered (either by importance or by logical time sequence). Pocock et al (2012) give a number of examples of the use of their technique using data from already published, completed trials which used hazard ratios instead of win ratios. The differences found help point to the extent to which treating all components as equally important affects interpretation. More experience with, and technical development of this technique will be necessary before its proper place in the statistical toolbox can be determined, but the general approach seems to have considerable merit.

This section has shown that it is possible to use, analyze, and interpret CEPs whose components have differing clinical importance. Should the problem under consideration dictate the use of a CEP with components of approximate equal weights, many of the potential disadvantages of CEPs disappear; but even when this is not possible, one can, with care, use heterogeneous CEPs.

6. Discussion

My thinking about clinical trial design considerations is guided by two complimentary quotations:

Sackett and Wennberg (1997, 1636) "the question being asked determines the appropriate research architecture, strategy, and tactics to be used – not tradition, authority, experts, paradigms, or schools of thought."

and

Lambert and Wood (2000, 164) "The starting point for experimental design should always involve joint consideration of the aims of the investigators and the constraints of the situation."

When considering phase III trials of new drug products, I have argued elsewhere (Kowalski et al 2008, 2012) that the only directly relevant outcome variables are length of life and QoL; while the use of surrogate outcomes may be of value in earlier phase trials bent on efficacy, when it comes time to judge what should be used in everyday clinical practice, there are no substitutes for survival and measures of what cost increased survival entails. The FDA's attitude toward the value of survival as an outcome in oncology drug testing was made clear by Johnson et al (2003, 1410): "An improvement in overall survival is the gold standard end point for a new oncology drug. The importance of a clinically meaningful survival improvement is unquestioned. Survival can be assessed with 100% accuracy for the event and with nearly 100% accuracy for the time of the event." Nevertheless, some flexibility in the choice of outcome has been documented. According to Johnson et al (2003, 1410), "Despite its importance, 68% (39 of 57) of the regular approvals and all of the 14 AAs [Accelerated Approvals] for oncology drugs were based on end points other than survival (January 1, 1990 to November 1, 2002)." That some of this flexibility is dependent upon study aims is evident from the work of Hirschfeld and Pazdur (2002, 139): "Patient benefit in oncology that would lead to full approval has been generally been [sic] applied as improved survival (primarily for cytotoxic medications and in particular for first line therapy), prolongation in time recurrence or disease-free survival (primarily in adjuvant trials), prolongation in time to progression (principally for hormonal or biological products), or palliation of symptoms, usually coupled with demonstration of objective tumor response ... In all cases there should not be a negative impact on survival." Thus, outcomes may be tailored to fit the measurement of just what the trial is intended to accomplish, provided only that one monitors mortality. The following example places an exclamation point on the importance of keeping an eye on mortality: Even if a trial's aim is to show that an intervention decreases the risk of non-fatal myocardial infarctions, Fleiss et al (1990, 685) noted, "Nonfatal myocardial infarction should not be used by itself as an end point. After all, it is the infarction that one is seeking to prevent, not just the nonfatal one."

Chi (2005, 610) also noted the primacy of mortality, but allowed that "In most cancer trials, improvement in overall survival is of primary clinical interest. However, relapse-free survival, or disease-free survival has been used in the adjuvant setting (after surgery) as a primary endpoint. Relapse-free survival can be viewed as a time to event composite endpoint (similarly for disease-free survival)

where an event is defined as either a relapse or death whichever comes first." He also thought that some additional measures of this kind might prove useful: "Progression-free survival can also be viewed as a time to event composite endpoint, where an event is either disease progression or death. Progression-free survival is gaining acceptance as a primary endpoint because crossover after progression may potentially reduce any subsequent treatment effect that might be observed on overall survival.[7] Time to treatment failure may also be viewed as a composite endpoint, where failure is defined as death, progression, or treatment withdrawal/crossover due to either lack of efficacy or toxicity." It needs to be recognized that the FDA may or may not view these endpoints as valid (able to support labeling claims), depending on the structure of the trial. For example, Pazdur (2008, 20-21), Director of the Office of Oncology Drug Products within the FDA, thought that time to treatment failure would seldom pass muster: "Time to treatment failure (TTF) is defined as the time from randomization to treatment discontinuation for any reason, including disease progression, treatment toxicity, patient preference, or death. From a regulatory point of view, TTF is generally not accepted as a valid endpoint. TTF is a composite endpoint influenced by factors unrelated to efficacy." I believe this to be a particularly good example of the Agency's preoccupation with *efficacy*, which falls short of the requirement of *effectiveness* that phase III trials should possess (Kowalski 2010), and allows such questionable practices as basing marketing approval decisions on trials utilizing placebo controls (Kowalski 2013). One might think that when judging how effective a given drug will be in everyday clinical practice – in the *real world*, if you will – it would be useful to see how it works *in that world*; in a world where patients can and do make treatment decisions using a wide variety of decision making strategies, including simply adopting (unsupported) preferences (Kowalski and Mrdjenovich 2013).

It is seen that, in many situations, the use of CEPs will remain contentious. There is one circumstance in which increasing event rates is seen less as a ruse to minimize expenditures, but more a product of the mother of necessity. As noted above, design questions must always be answered with due regard for the "constraints of the situation." A constraint that may impact on the choice of outcome, especially in oncology trials, is small numbers of even *potential* subjects available for participation. Cancer patients are often reluctant to participate in clinical trials, and this problem is compounded in the case of a rare cancer where few individuals will even be potentially available for study. It may be that in many such cases an outcome like relapse-free survival (with associated CEP = death + relapse) will make sense as a study aim *and* provide more input for data analysis.

[7] Progression-free survival is recorded before patients change therapies, so the results cannot be obscured by subsequent cross-over therapies.

One need only to consider the (purely) explanatory to (purely) pragmatic spectrum of clinical trials to realize that it will be difficult to provide useful rules for design questions that will cover all the intermediate cases. I will recognize two ends of another continuum for which the use of CEPs should be at least potentially useful. On the one hand, Johnson et al (2003, 1408) thought, again echoing the FDA's viewpoint, "A composite end point may be appropriate when the drug's benefit is multifaceted. The end-point components should be related and generally of similar clinical importance." When this is the case, it will not be necessary to pursue weighted/hierarchical approaches to CEP construction, but interpretation may have to be judiciously circumscribed. Ferreira-Gonzàlez et al (2008, 286) "the sponsor of a trial with a particular drug may prefer to focus on a positive result based on a CEP rather than to enter into debate about the precaution needed in the interpretation of the treatment effect." Even should the FDA insist that labeling should be transparent about what the drug can and cannot be expected to accomplish, precautions regarding interpretation are apt to be lost in the small print. Chi (2005) gives some examples where drug labeling includes reference to (only) significance of CEPs, but where there is ample opportunity to extrapolate benefit to the components without appropriate statistical documentation of these effects.

On the other hand, a drug may target a particular manifestation of a disease process, e.g., to reduce non-fatal AMIs. In this case, as we have seen, non-fatal AMIs can be a useful outcome, as long as it is joined by all-cause mortality. Some consider that cause-specific mortality might be more sensitive to certain treatment difference (e.g., Yusuf and Negassa 2002), but Chi (2005, 613) urged caution: "The use of cause-specific mortality may introduce informative censoring in the analysis of the composite endpoint. For instance, patients who die of treatment-related toxicity will be considered as censored cases. In the analysis of the composite endpoint, this may introduce a bias favoring the treatment. It is recommended that all-cause mortality should be used." Another reason for avoiding cause-specific mortality is the difficulty with which cause can be reliably established. Lauer and Topol (2003, 2575) thought that, "Among fatal end points, only all-cause mortality can be considered objective, unbiased, and clinically relevant. As previously reviewed in depth, the use of end points such as 'cardiac death,' 'vascular death,' and 'arrhythmic death' are inherently subject to error due to biased assessment and to the biological complexities of disease, especially among elderly individuals." See also Gottlieb (1997) and Lauer et al (1999).

Whether or not it will be sensible to use a CEP in a given situation depends on the situation. Saving time and money has much to recommend it in many of life's endeavors, but not when one is investigating the safety and effectiveness of new drug products. Conservation of resources may be required by practical constraints, as in the case of a rare cancer where not enough patients even exist to populate a

"traditional trial," but CEP use in such cases is an attempt to maximize information, not profits. As should be evident from the above discussion, it is difficult to prescribe the use of CEPs except to say "when indicated by the question the trial is being designed to answer." The only other general rule I would recommend is that (virtually all) CEPs should include all-cause mortality among their components. We want the drugs we take to increase the length and quality of our lives, and the only way to fairly judge whether or not they will do this is to try them out and measure these outcomes. We will want to be able to see not only whether drugs promote longevity, but also to check that drugs that favorably impact QoL do not do so by imposing an unacceptable risk of an early demise. I would also think that event-free survival type outcomes will often be the answer to clinically interesting questions.[8] Since these can be seen as CEPs with mortality and the event(s) of interest as components, when event-free survival does indeed answer the question posed, an appropriately structured CEP is indicated. In time to treatment failure (TTF) studies, where one seeks to determine the causes of treatment failure and to assess the extent to which each cause contributes to the total failure rate, this approach can be especially rewarding. For a good example, see Arriagada et al (1992). Note that they use a statistical technique based on a competing risk approach that avoids biases related to assumptions of independence among the events incurred by the conventional Kaplan-Meier or actuarial methods.

I want to close this discussion with an example, illustrating just one of the myriad ways in which constructing a composite *strictly to save time/money* may cause more problems than it was intended to solve. A clinical trial steering committee, using information supplied to it by the trial's data and safety monitoring board (DSMB), actually changed the choice of the primary outcome variable during the course of a trial. According to Freemantle et al (2003), this occurred during the so-called CAPRICON trial that was investigating the effects of a beta-blocker on patients with left ventricular dysfunction following myocardial infarction. The original primary outcome was all-cause mortality, but when too few of the subjects cooperated by expiring in a timely fashion, the decision was made to add cardiovascular hospital admissions to deaths in order to increase the event rate and so keep up with the study timetable. The 5% available for the type I error rate was split into 0.045 for the new composite (CEP = cardiovascular hospital admissions + deaths) and 0.005 for the now secondary outcome of all-cause mortality. As luck would have it, when this altered trial was completed, the P-value for the CEP was 0.30, while for the secondary outcome, P = 0.03. Neither the primary nor the secondary outcomes were significant after the alpha-splitting technique was employed; but P = 0.03 would have been significant under the original design with all-cause mortality primary (P = 0.03 < 0.05). In addition to pointing to some of the questions that can arise when DSMBs are employed (Kowalski and Hewett 2009), this example shows that when using the CEP, the treatment effect of most interest was diluted by the addition of an outcome that exhibited no effect.

[8] As a rough indication of this, searching in Google Scholar for *event-free survival* got 54,100 results. Searching on *time to treatment failure* produced 2,070,000.

References

[1] Arriagada, R., L.E. Rutqvist, A. Kramar, and H. Johansson. 1992. Competing risks determining event-free survival in early breast cancer. British J Cancer 66: 951-7.

[2] Berry, D.A., and Y. Hochberg. 1999. Bayesian perspectives on multiple comparisons. J Statistical Planning and Inference 82: 215-27.

[3] Braunwald E., C.P. Cannon, C.H. McCabe. 1992. An approach to evaluating thrombolytic therapy in acute myocardial infarction: the unsatisfactory outcome endpoint. Circulation 86: 683-7.

[4] Cannon, C.P. 1997. Clinical perspectives on the use of composite endpoints. Controlled Clinical Trials 18: 517-29.

[5] Chi, G.Y.H. 2005. Some issues with composite endpoints in clinical trials. Fundamental & Clinical Pharmacology 19: 609-19.

[6] FDA. 2006. Guidance for Industry. Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. Available at: http://www.fda.gov/cder/guidance/index.htm

[7] Ferreira-Gonzàlez, I., P. Alonso-Coello, I. Solà, et al. 2008. Composite endpoints in clinical trials. *Revista Española de Cardiologia* 61: 283-90.

[8] Ferreira-Gonzàlez, I., G. Permanyer-Miralda, J.W. Busse, et al. 2007. Methodologic discussions for using and interpreting composite endpoints are limited, but still identify major concerns. *J Clinical Epidemiology* 60: 651-7.

[9] Ferreira-Gonzàlez, I., G. Permanyer-Miralda, J.W. Busse, et al. 2007. Composite endpoints in clinical trials: The trees and the forest. J Clinical Epidemiology 60: 660-1.

[10] Fleiss, J.L., T. Bigger, Jr., M. McDermott, et al. 1990. Nonfatal myocardial infarction is, by itself, an inappropriate end point in clinical trials in cardiology. Circulation 81: 684-5.

[11] Freemantle, N. 2001. Interpreting the results of secondary endpoints and subgroup analyses in clinical trials: Should we lock the crazy aunt in the attic? BMJ 322: 989-91.

[12] Freemantle, N., and M. Calvert. 2007. Composite and surrogate outcomes in randomized controlled trials: Composite end points may mislead – and regulators allow it to happen. *BMJ* 334: 756-7.

[13] Freemantle, N., M. Calvert, J. Wood, J. Eastaugh, and C. Griffin. 2003. Composite outcomes in clinical trials: Greater precision but with greater uncertainty. JAMA 289: 2554-9.

[14] Friedman, L.M., C.D. Furberg, and D.L. DeMets. 1998. Fundamentals of Clinical Trials, third edition. New York:

Springer.

[15] Gottlieb, S.S. 1997. Dead is dead – artificial definitions are no substitute. Lancet 349: 662-3.

[16] Hirschfeld, S., and R. Pazdur. 2002. Oncology drug development: United States Food and Drug Administration perspective. Critical Reviews in Oncology/Hematology 42: 137-43.

[17] ICH. 1999. International Conference on Harmonization of Technical Requirements for Human Use. ICH harmonized tripartite guideline: statistical principles for clinical trials. Statistics in Medicine 18: 1905-42.

[18] Johnson, J.R., G. Williams, and R. Pazdur.2003. End points and United States Food and Drug Administration approval of oncology drugs. *J Clinical Oncology* 21: 1404-11.

[19] Kowalski, C.J. 1972. A commentary on the use of multivariate statistical methods in anthropometric research. *Am J Physical Anthropology* 36: 119-31.

[20] Kowalski, C.J. 2010. Pragmatic problems with clinical equipoise. Perspectives in Biology and Medicine 53: 161-73.

[21] Kowalski, C.J. 2013. Clinical trials of new drug products: What gets compared to whom? Perspectives in Biology and Medicine In press.

[22] Kowalski, C.J., J. Bernheim, N.A. Birk, and P. Theuns. 2012. Felicitometric hermeneutics: Interpreting quality of life measurements. Theoretical Medicine and Bioethics 33: 207-20.

[23] Kowalski, C.J., and J.L. Hewett. 2009. Data and safety monitoring boards: Some enduring questions. J Law, Medicine & Ethics 37: 496-506.

[24] Kowalski, C.J., and A. Mrdjenovich. 2013. Patient preference clinical trials: Why and when they will sometimes be preferred. Perspectives in Biology and Medicine 56: 18-35.

[25] Kowalski, C.J., S. Pennell, and A. Vinokur. 2008. Felicitometry: Measuring the 'quality' in quality of life. Bioethics *22:* 307-13.

[26] Lambert, M.F., and J. Wood. 2000. Incorporating patient preferences into randomized trials. J Clinical Epidemiology 53: 163-6.

[27] Lauer, M.S., E.H. Blackstone, J.B. Young, and E.J. Topol. 1999. Cause of death in clinical research: time for a reassessment? J Am Coll Cardiology 34: 618-20.

[28] Lauer, M.S., and E.J. Topol. 2003. Clinical trials – multiple treatments, multiple end points, and multiple lessons. JAMA 2003: 2575-7.

[29] Lubsen, J., and B-A. Kirwan. 2002. Combined endpoints: can we use them? Statistics in Medicine 21: 2959-70.

[30] Montori, V.M., G. Permanyer-Miralda, I. Ferreira-Gonzàlez, J.W. Busse, et al. 2005. Validity of composite end points in clinical trials. BMJ 330: 594-6.

[31] Neaton, J.D., G. Gray, B.D. Zuckerman, and M.A. Konstam. 2005. Key issues in end point selection for heart failure trials: Composite end points. J Cardiac Failure 11: 567-75.

[32] O'Brien, P.C., and N.L. Geller. 1997. Interpreting tests for efficacy in clinical trials with multiple endpoints. Controlled Clinical Trials 18: 222-7.

[33] Pazdur, R. 2008. Endpoints for assessing drug activity in clinical trials. The Oncologist 13 (suppl 2): 19-21.

[34] Pocock, S.J., C.A. Ariti, T.J. Collier, and D. Wang. 2012. The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. European Heart Journal 33: 176-82.

[35] Rothman, K.J. 1990. No adjustments are needed for multiple comparisons. Epidemiology 1: 43-6.

[36] Sackett, D.L., and J.E. Wennberg. 1997. Choosing the best research design for each question: It's time to stop squabbling over the "best" methods. BMJ 315: 1636.

[37] Schultz, K.F., and D.A. Grimes. 2005. Multiplicity in randomized trials I: endpoints and treatments. Lancet 365: 1591-5.

[38] Schwartz, D., R. Flamant, and J. Lellouch. Translated by M.J.R. Healy. 1980. Clinical Trials. New York: Academic Press.

[39] Torp-Pedersen, C., M. Móller, P.E. Bloch-Thomsen, L. Kober et al. 1999. Dofetilide in patients with congestive heart failure and left ventricular dysfunction. NEJM 341: 857-65.

[40] Yusuf, S., and A. Negassa. 2002. Choice of clinical outcomes in randomized trials of heart failure therapies: disease-specific or overall outcomes? Am Heart J 143: 22-8.

[41] Zhang, J., H. Quan, J. Ng, and M.E. Stepanavage. 1997. Some statistical methods for multiple endpoints in clinical trials. Controlled Clinical Trials 18: 204-21.

[42] Zivin, J.A. 2000. Understanding clinical trials. Scientific American 282: 69-75.