

## Perspective

# Key Issues in End Point Selection for Heart Failure Trials: Composite End Points

JAMES D. NEATON, PhD,<sup>1</sup> GERRY GRAY, PhD,<sup>2</sup> BRAM D. ZUCKERMAN, MD,<sup>2</sup> AND MARVIN A. KONSTAM, MD<sup>3</sup>

*Minneapolis, Minnesota; Rockville, Maryland; Boston, Massachusetts*

---

## ABSTRACT

**Background:** Composite outcomes are commonly used in heart failure trials. The aim of this article is to discuss the advantages and disadvantages of composite outcomes and recommend guidelines for reporting them. Examples are used to illustrate key points.

**Methods and Results:** A workshop jointly planned by the Heart Failure Society of America and the US Food and Drug Administration was convened in April 2004. One of the panel discussions concerned the use of composite outcomes in heart failure trials. With use of composite outcomes, event rates are higher and if it is reasonable to assume that the treatment effect is similar for each component of the composite outcome, sample size will be smaller than using one of the components as the primary end point. Composite end points are difficult to interpret if effects are not similar for all components or if the effect of treatment is primarily on a more common, less serious component of the composite. Composite outcomes typically only focus on the first occurring event. This can lead to a substantial loss of information in some trials. When composite end points are used, data collection for all components should continue until the end of the trial so that each component can be separately evaluated.

**Conclusion:** Composite end points should be used with caution. Additional research is needed on optimally weighting components of composite outcomes and to better using the entire event history of patients in heart failure trials. Further analyses of completed trials could be useful in this respect.

**Key Words:** Composite end point, clinical trial, heart failure, medical devices.

---

The selection of a primary end point is an important step in a clinical trial. Along with the treatments and the definition of the target population, the primary end point defines the research question. Many trials use composite outcomes

as primary or secondary outcomes. Meinert defined a composite outcome as “an event that is considered to have occurred if any one of several different events or outcomes is observed.”<sup>1</sup> The terms *composite end point* and *combined end point* are used interchangeably by many clinical trialists. Use of composites is a common way to deal with multiple outcomes.

The purpose of this article is to discuss the use of composite outcomes in heart failure trials, give recommendations for the reporting of composite outcomes, and discuss some advantages and disadvantages of a single versus composite outcome versus other approaches for handling multiple outcomes. Examples are cited to illustrate how composites have been defined and used in heart failure drug and device trials.

## General End Point Considerations

The objectives and scope of clinical trials are largely determined by the end point chosen. For example, a 24-week

---

*From the*<sup>1</sup>*Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota;*<sup>2</sup>*Center for Devices and Radiological Health, US Food and Drug Administration, Rockville, Maryland;*<sup>3</sup>*Division of Cardiology, Tuft-New England Medical Center, Boston, Massachusetts.*

Reprint requests: James D. Neaton, PhD, Division of Biostatistics, School of Public Health, University of Minnesota, 2221 University Ave SE, Room 200, Minneapolis, MN 55414.

The opinions expressed in this paper are those of the authors and do not necessarily represent those of the editor or the Heart Failure Society of America (HFSA). All writing group members were required to complete and submit a Faculty Disclosure Questionnaire before the workshop was held, April 1–2, 2004. The workshop was sponsored by the HFSA. This article represents the professional opinion of the authors and is not an official document, agency guidance, or policy of the U.S. Government, the Department of Health and Human Services, or the Food and Drug Administration, nor should any official endorsement be inferred.

1071-9164/\$ - see front matter

© 2005 Elsevier Inc. All rights reserved.

doi:10.1016/j.cardfail.2005.08.350

clinical trial might investigate the effect of a cardiac resynchronization device on quality of life, but it could not reliably assess the influence of the device on morbidity and mortality. For the latter, a larger study with longer follow-up would be needed. Both types of studies are important for a full understanding of the efficacy and safety of a device.

General factors to consider in choosing a primary end point are discussed in several clinical trial texts.<sup>2-4</sup> 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.

Other trial conditions, such as blinding, often dictate the choice of the primary outcome. For example, change in patient-reported quality of life or in investigator-assessed functional status after a period of treatment may be reasonable choices for an outcome in a double-blind study of a new heart failure treatment. However, in a non-blind study, as many device trials are, these outcomes may not be ideal as primary end points because of the potential for bias.

End points may be categorized as (1) measures of clinical outcomes (eg, death or morbid events), (2) measures of symptoms or clinical status (eg, quality of life or New York Heart Association [NYHA] class change, or (3) surrogates (eg, ejection fraction, ventricular volumes, B-type natriuretic peptide). Surrogate end points are those that are not direct measures of clinical outcome, symptoms, or clinical status, but correlate with clinically relevant findings, either because they signal worsening of the underlying disease or contribute to its pathophysiology. Importantly, as noted by Fleming and DeMets “a correlate does not a surrogate make.”<sup>5</sup> Prentice argued that a valid surrogate must fully capture the net effect of treatment on the clinical outcome.<sup>6</sup> There are no established surrogates for device trials. There may be “partial” surrogates. Further research is needed in this area. Pooling of data from trials that measured both surrogate markers and clinical outcomes might identify potential surrogates. There is a precedent for this for anti-retroviral drugs used to treat HIV.<sup>7</sup>

Treatment differences based on clinical outcomes and on surrogate markers are not always consistent. Fleming and DeMets<sup>5</sup> cite the PROMISE (Prospective Milrinone Survival Evaluation) trial, which found that milrinone, a drug treatment that increased exercise tolerance, increased total mortality in patients with advanced heart failure.<sup>8</sup> Knowledge or suspicion of an increase in mortality would represent a major obstacle to approval of a drug or device, despite known improvement in quality of life or functional capacity. However, such approval may be warranted under some circumstances, and measures that integrate survival and quality of life should be considered.

Use of surrogate end points in clinical trials offer many potential advantages: fewer patients, shorter follow-up, and lower cost.<sup>9</sup> However, use of surrogates requires a clear understanding of the relationship—both physiologic and

statistical—between the surrogate and the clinical results that are presumed to follow. Demonstration of “efficacy” based on surrogate results must be further subjected to analysis of risk-benefit, because serious adverse events may negate an intervention’s clinical utility. An appropriate approach may be to integrate surrogate end points with clinical measures through use of composites, allowing the surrogate finding to augment the clinical outcome, which might otherwise not be definitive on its own.

### Rationale for Using Composite End Points

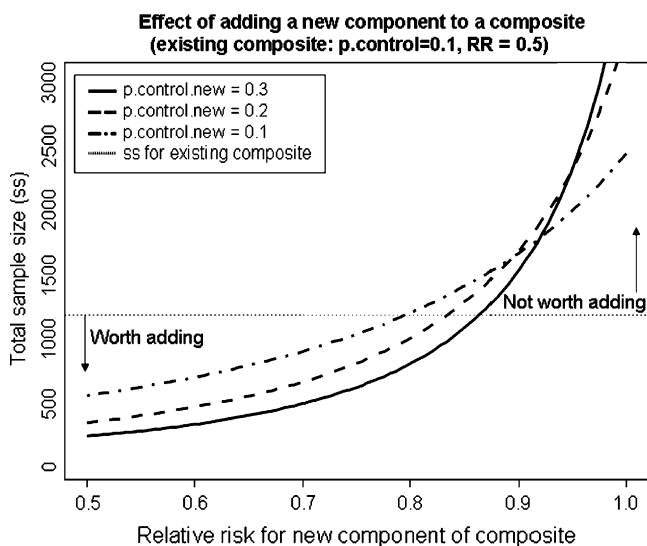
Some of the characteristics of surrogate outcomes also motivate the use of composite end points. Composite end points may involve clinical outcomes, surrogate outcomes, or combinations of both. Each component of a composite outcome must be ascertainable without bias, and each component should be clinically relevant and sensitive to the hypothesized effects of the treatment.

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. Yusuf and Negassa<sup>10</sup> discuss this point in the choice of end points for heart failure trials and argue for cause-specific end points (cardiovascular disease [CVD] mortality and CVD hospitalization) rather than all-cause mortality and hospitalization because they are more sensitive to treatment differences and power is increased. The optimal degree of specificity will depend on the target population. For example, in a target population that had just experienced an acute myocardial infarction such as enrolled in the Eplerenone Post-Acute Myocardial Infarction Heart Failure Efficacy and Survival Study (EPHESUS) heart-failure study, the majority of deaths would be expected to be CVD (86% were).<sup>11</sup> Thus using all-cause mortality instead of CVD would not be expected to result in a very large loss of sensitivity. In contrast, in a study of patients with stable disease and preserved left ventricular ejection fraction, more patients might be expected to die from non-CVD causes than in study of patients with acute MI. Thus use of an all-cause mortality outcome might result in a substantial loss of power for an intervention aimed at preventing CVD. For example, in the Candesartan in Heart Failure Assessment of Reduction in Mortality (CHARM)-Preserved Trial, 71% of deaths were attributed to CVD disease.<sup>12</sup>

This same thinking applies to the construction of composites. If the expected risk reduction or hazard ratio is a priori considered similar for all components of the end point, sample size will usually be less than using any one of the components. For example, if the estimated failure rate in the control group is 10%, the required sample size to detect

a 50% lower rate (5%) in a group given a new device is 1170 patients (585 per group) assuming a significance level of 0.05 (2-sided) and power of 0.90. If, by using a composite outcome, the control failure rate was increased to 30%, the sample size required to detect a 50% reduction in this rate (to 15%) at the same type I and II error levels is substantially smaller—330 patients (165 per group). However, if by using the composite end point, the expected reduction in risk is reduced to 25% instead of 50% (an expected rate of 22.5% for the device group instead of 15%), the sample size required would be 1450 (725 per group). This is shown more generally in Fig. 1. The horizontal line corresponds to the sample size for the 10% control event rate and a 50% lower rate in the new device group. The 3 curves illustrate the impact on sample size of adding new component(s) to the composite. The 3 curves correspond to new component(s) that increase the overall composite event rate by 10%, 20%, and 30%, respectively. The middle curve, corresponding to a new component that adds 20%, shows that the expected risk reduction for the new component(s) need to be nearly 20% to realize any sample size efficiency for the higher overall control event rate.

Another reason that composites are used is to avoid the problem of competing risks. For example, in a trial of patients with advanced heart failure, an end point of hospitalization for heart failure would be criticized because it does not account for mortality. The censoring of deaths in such a trial (patients who died without hospitalization) is probably “informative” because a patient who is censored for death is likely not at the same risk of hospitalization (had they survived) as a patient who survived as long and is still at risk for hospitalization. If censoring because of death, or reasons for it, varied by treatment group, the estimate of the treatment effect would be biased. Related to this, if the device appeared to reduce the rate of hospitalization but increased mortality, the results would be



**Fig. 1.** Impact on sample size of adding new components to a composite income.

difficult to interpret. Similarly, an end point of change in self-reported quality of life after 24 weeks would also be suspect if deaths or serious illness prevented a quality of life assessment from being carried out and this was not accounted for in the primary analysis. The patients without the quality of life assessment may not be representative of those with it.

As noted previously, a primary end point needs to be ascertainable for all, or at least nearly all, of the patients enrolled. One way to account for missing data from mortality in such a trial is to include death as part of a composite end point. Califf and colleagues<sup>13</sup> made this argument concerning use of left ventricular ejection fraction in thrombolytic therapy trials. He noted that a problem with use of such an end point is that data might not be collected because the patient was too sick or had died. Use of a composite ordinal outcome that included mortality as the worst outcome was proposed as a solution.

A third reason composite end points are used was cited by Freemantle and colleagues.<sup>14</sup> He quoted a guideline by the International Conference on Harmonization that states: “If a single primary variable cannot be selected..., another useful strategy is to integrate or combine the multiple measurements into a single or composite variable.”<sup>15</sup> Inability to reach consensus on a single outcome is generally not a good reason to use a composite end point.

### Cautions with Use of Composite End Points

One caution, already mentioned, is that there could be a loss of power if the treatment effect is not similar for all of the components. Worse yet is the situation when components of the combined end point go in opposite directions. If this occurs, findings of the study could be ambiguous, and clear recommendations about the use of the treatment might not be possible. This problem could manifest itself if the earlier occurring outcomes were less likely to represent a major failure of the treatment compared with later occurring outcomes. According to the definition given previously, if any component of a composite failure end point occurs, the patient is considered a “failure.” In time to event analyses, the failure time corresponds to the time from randomization to the first occurrence of any event included in the composite. Many treatments have modest (not large) effects on progression of heart failure. Thus it is possible that an effective treatment for patients with advanced heart failure might reduce risk of death but not hospitalization from any cause. A first event analysis of a composite defined as all-cause mortality or all-cause hospitalization might not be sensitive to the effects of a treatment with modest efficacy.

Another caution with the use of composite end points relates to the weighting of individual components. In most studies, the components of the composite end point are assigned equal weight even though patients and clinicians may not consider them equally important. For example, the composite outcome of mortality or hospitalization from

any cause as mentioned in the preceding paragraph assigns equal weight to a death from CVD and a hospitalization for an appendectomy.

An alternative to equal weighting is to assign each component a weight that reflects relative severity to other components. Assignment of weights to components may be based on subjective rankings or objective criteria. To obtain a weighting for different clinical outcomes, Califf and colleagues surveyed 407 cardiologists at the 1989 American Heart Association meeting concerning end points that might be affected by a reperfusion strategy.<sup>13</sup> The cardiologists were asked to rank outcomes on a scale from 0 to 10, with 10 being the most serious. They proposed using the ranking of the outcomes experienced by a patients to define a single, ordered end point.

If weights are assigned to the individual components, it is important that the weights be validated by relating the weighted composite to some credible outcome (eg, mortality). This might be accomplished by applying proposed weighting schemes to completed clinical trials in which the various end points were assessed but not weighted in the primary analysis. Neaton and colleagues<sup>16</sup> uses the results of completed HIV treatment trials and compares the use of subjective rankings by patients and clinicians with rankings based on risk of death as a means for weighting opportunistic illnesses associated with HIV. Bjorling and Hodges study different rule-based ranking schemes of event histories and validated them using expert rankings.<sup>17</sup> These findings are relevant to trials in many areas including heart failure. Not surprisingly, ranking schemes in best accord with expert opinion considered the severity of different events, the timing events, and how many events each patient experienced.

Another caution, related to the discussion previously, is if the treatment benefit with the composite results primarily from a common, less serious outcome, and the less common, more serious components of the composite do not differ between the treatment groups or go in the other direction.

Montori and colleagues<sup>18</sup> summarize these cautions by stating 3 questions to consider as a guide in the choice of a composite outcome: (1) Are the components of the composite end point of similar importance to patients? (2) Did the more and less important outcomes occur with similar frequency? (3) Are the components likely to be similarly effected by the treatment?

Because of these concerns with the use of composites, it is important to summarize each component of the composite separately in the study report. This is discussed further in the following section.

### Reporting of Composite Outcomes

In trials with a composite end point, patients should be followed to the end of the study for all components of the composite outcome. Continued data collection of all outcomes comprising the composite end point will permit a proper intent-to-treat analysis of each component of the

composite as well as the composite (Analysis 1). This has been referred to as the “Consumer Reports” analysis.<sup>13</sup> For example, if the composite outcome is cardiovascular mortality, nonfatal myocardial infarction or nonfatal stroke, patients who experience a nonfatal stroke should continue to be followed for a nonfatal myocardial infarction or cardiovascular death.

The trial report should also describe the frequency with which each component occurred as the first event for those with end points (Analysis 2). For example, if 100 patients assigned the device experienced the composite outcome described previously (cardiovascular mortality, nonfatal myocardial infarction, or nonfatal stroke), the trial report should state how many of these 100 events were a result of cardiovascular mortality, and how many were nonfatal myocardial infarction or stroke.

These 2 analyses will allow the reader (the consumer of your trial report) to assess whether treatment differences for each component of the composite trend in the same direction (Analysis 1), and whether the first event experienced by those with an end point is similar for the treatment groups (Analysis 2). Freemantle and colleagues<sup>14</sup> emphasize the importance of Analysis 1 in trial reports. In addition, Freemantle and colleagues advocate defining each component of the composite end point as a secondary outcome. A recent report of a trial of nifedipine illustrates proper reporting of composite outcomes<sup>19</sup> (see Tables 3 and 4 in that article for illustrations of Analysis 1 and Analysis 2).

More generally, in device trials, all outcomes, both safety and efficacy, should be collected through the end of the study. This will minimize the bias in assessing the relative benefits and risk that might otherwise arise due to censoring of some outcomes and then having to carry out an analysis that is not intent to treat.

Collection of all outcomes, even repeat occurrences of the same event (eg, multiple nonfatal myocardial infarctions in a study like that described previously), might also improve the precision of some analyses that take into account the entire event profile of the patient while in the study. This is discussed further in a later section.

### Examples of Heart Failure Trials With Composite Outcomes

The use of composite outcomes is very common in trials of new drugs and devices for patients with heart failure. A few examples are cited to illustrate their use.

#### EPHESUS Trial of Eplerenone

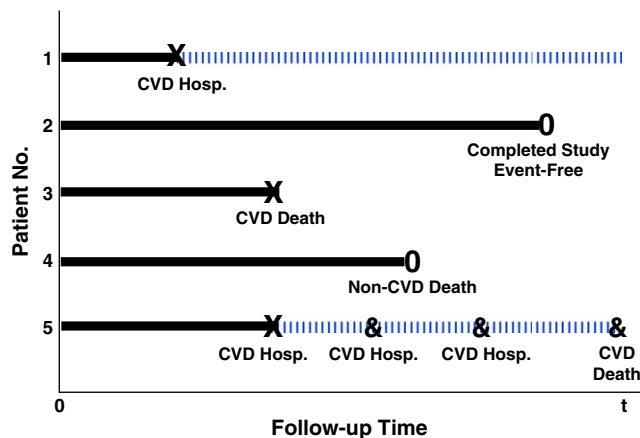
In the EPHESUS trial, 2 primary end points were defined.<sup>11</sup> One of the coprimary end points was death from any cause; the second coprimary end point was death from CVD or first hospitalization for a CVD event—heart failure, recurrent myocardial infarction (patients had a documented myocardial infarction at entry), stroke, or ventricular arrhythmia. The use of coprimary end points

will be discussed in a later section. For now, consider Fig. 1, which illustrates how patients were counted for the composite end point of CVD death or hospitalization that was used in EPHEBUS.

The event history of 5 patients is illustrated in Fig. 2. Time zero is randomization. The “X” and “O” denote event and censoring times, respectively, for the time to event analysis. The “&” denotes events that do not count for the composite end point analysis. The first patient was hospitalized for CVD shortly after randomization; the second patient did not experience the composite outcome (was event-free at the end of the trial); the third patient died from CVD, which was not preceded by a CVD hospitalization; the fourth patient died from a non-CVD cause and is censored at that time (the patient is no longer at risk for CVD death); the fifth patient was hospitalized for CVD 3 times before dying from CVD (only the first event is counted for the composite end point).

These fictitious patients illustrate some characteristics of composite end points analyzed with time to event methods: (1) although a CVD death would generally be regarded as more serious than a CVD hospitalization, the shortest failure time is for the first patient who was hospitalized for CVD; (2) because non-CVD deaths are treated as censored observations and do not count as events, a priori, it would be important to expect no effect of treatment on non-CVD mortality because if there were, the interpretation of the composite could be complicated; and (3) even though the fifth patient had the more severe event history, only the first event is counted and it occurred after the primary events occurred for the first and third patients.

In the EPHEBUS trial, a “Consumer Reports” analysis as described previously was cited for the composite end point. For the composite, 885 of 3319 patients assigned eplerenone and 993 of 3313 patients assigned placebo experienced at least 1 event. Deaths from CVD numbered 407 and 483 for eplerenone and placebo, respectively; CVD hospitalizations numbered 606 and 649, respectively. Thus



**Fig. 2.** Illustrative patients in a time-to-event trial with a composite outcome of cardiovascular (CVD) hospitalization or death from CVD.

128 patients in the eplerenone group and 139 patients in the placebo group were hospitalized at least once and died from CVD. Hazard ratios for the composite end point, for CVD mortality, and for CVD hospitalization were 0.87, 0.83, and 0.91, respectively.

CVD deaths and hospitalizations were further broken down by cause. Because all hospitalizations were collected until the end of the study, it was also possible to compute hazard ratios for type of nonfatal CVD event (see Table 2 of Pitt and colleagues).

### COMPANION Trial

The COMPANION (ie, Comparison of Medical Therapy, Pacing, and Defibrillation in Heart Failure) trial evaluated pacemakers with and without a defibrillator versus optimal pharmacologic therapy alone in patients with advanced heart failure. The primary end point of the COMPANION trial was the composite of death from any cause or hospitalization from any cause.<sup>20</sup>

In this trial, 26% of patients in the pharmacologic therapy group withdrew from the study to receive a commercially available pacemaker implant because of arrhythmia or heart failure. Although about half of these patients had experienced a hospitalization event, mortality follow-up was not available. In contrast, the withdrawal rates for the pacemaker and pacemaker-defibrillator arms were much lower—6% and 7%, respectively. Fortunately the sponsor attempted to reconsent these patients for event data collection through the closing date of the study. As a consequence of those efforts, the primary end point was unknown for 9% of the control patients and vital status was unknown for 4%. The corresponding percents for both device arms were 1%.

This trial illustrates the importance of planning for data collection of all events for all randomized patients from the outset. This is particularly important for survival.

### CardioWest Artificial Heart Study

Composites are used for success/failure outcomes as well as time to event outcomes. For example, a nonrandomized study with historical controls was carried out to evaluate the safety and efficacy of the CardioWest Total Artificial Heart in transplant-eligible patients at risk for imminent death from irreversible biventricular cardiac failure.<sup>21</sup> One of the end points used to evaluate the artificial heart was a composite end point called “treatment success.” Treatment was considered successful if, 30 days after transplantation, the patient was: (1) alive; (2) in NYHA class I or II; (3) ambulatory; (4) not dependent on a ventilator; and (5) not undergoing dialysis. Some components of this composite are more easily ascertainable than others. With this type of outcome, the timing of events within the planned follow-up period is ignored.

### PICO Study

Composites can also be defined using both continuous outcomes and clinical events. The Pimobendan in

Congestive Heart Failure (ie, PICO) trial was a 24 week placebo-controlled study of the addition of pimobendan (versus placebo) on exercise duration.<sup>22</sup> The investigators observed that 24-week data on exercise duration, the primary end point, were missing for some patients because of death or because of a cardiovascular contraindication. Furthermore, the number and reason for missing the exercise duration data varied by treatment group—the missing data were likely informative. Thus they carried out a rank analysis in which deaths and missing data from cardiovascular contraindications were ranked lower than the lowest exercise time change observed.

Lubsen<sup>23</sup> discussed this example of a composite. More generally, a ranking of outcomes has been advocated by others<sup>13,24–26</sup> as means of creating an ordinal composite outcome that takes into account the relative importance of the different components.

### Gamma-One Study

The Gamma-One study is an example of a trial where more serious, less frequent components of a composite went in a direction opposite that to a more common, less severe component. This trial compared local intracoronary radiation (brachytherapy) versus “placebo” therapy in 252 patients with documented myocardial ischemia from in-stent stenosis. The brachytherapy group (n = 131) received an indwelling intracoronary ribbon containing a sealed source of iridium-192; the “placebo” group (n = 121) received a similar appearing nonradioactive ribbon. The primary end point was a composite that included death, myocardial infarction (including late thrombosis), emergency bypass surgery, and the need for revascularization of the target lesion after 9 months.<sup>27</sup> Deaths, myocardial infarctions, and late thrombosis favored the control arm (4 versus 1, 13 versus 5, and 7 versus 1, respectively); revascularizations of the target lesion favored the treatment group (32 versus 51). The composite also favored the treatment group (37 versus 53). The differential trends delayed Food and Drug Administration approval, but ultimately, the device was approved with a warning to avoid use with the placement of new stents (all of the thrombotic events occurred among patients who received additional stents at the time of brachytherapy).<sup>28</sup> This trial illustrates the importance of long-term follow-up to ensure that initial favorable effects are not later overwhelmed by more serious adverse effects.

### A-HeFT Study

In some situations it may be desirable to use a composite in which the components are defined over different periods. The African-American Heart Failure Trial (ie, A-HeFT) was a double-blind placebo controlled trial of isosorbide dinitrate and hydralazine versus placebo in patients with NYHA class III or IV heart failure. The primary end point of the study was a composite score that included clinical outcomes (death and hospitalization for heart failure) at any time during the study and change in quality of life after 6

months (all patients were to be followed for at least 6 months).<sup>29,30</sup> The investigators scored events as follows: (1) death (−3); (2) first hospitalization for heart failure (−1); and (3) change in quality of life at 6 months—by 10 or more units (+2), 5–9 units (+1), less than a 5-unit change (0), worsening by 5–9 units (−1), and worsening by 10 or more units (−2). Quality of life was measured using the Minnesota Living with Heart Failure Questionnaire, which has a score range from 0 (best) to 105 (worst). This gives a range of possible scores of −6 to +2 for the primary end point.

The study was recently stopped because of a significantly higher mortality rate in the placebo group compared with the group given isosorbide dinitrate plus hydralazine.<sup>30</sup> The overall score and each of the components favored the isosorbide dinitrate plus hydralazine group.<sup>30</sup> For example, the primary composite score averaged −0.1 (standard deviation [SD] = 1.9) for the isosorbide dinitrate plus hydralazine group and −0.5 (SD = 2.0) for the placebo group. Because estimates of the SD for change in this composite score were not available at the beginning of the study, sample reestimation was built into the study as part of the interim analyses (SD was initially assumed to be between 1 and 2). This is an important consideration of a new composite scoring system is used.

The composite scoring system used in A-HeFT results in a score for each patient. An advantage of the scoring system is that it integrates quality of life with clinical outcomes. A disadvantage is that the composite does not take account of the time of the hospitalization or death (eg, a death at 1 year was weighted the same as a death at 1 month). This could result in a loss of power if the composite was used in a longer term study in which the majority of patients were hospitalized or died. In A-HeFT, patients were not followed beyond 18 months. Another general disadvantage of such composite scores is that it may be difficult to achieve consensus on an appropriate weighting scheme when the components of the score are very different in terms of their seriousness.

### Alternatives to a Composite End Point

There are several alternatives to using a composite primary end point. A single outcome variable such as all-cause mortality could be specified. Although the cautions cited for a composite outcome do not apply, sample size and power considerations may not permit the use of such an outcome. Even with a single primary outcome variable, the collection and analysis of multiple end points is essential in clinical trials for a full understanding of the effects of a treatment. Pocock notes that a patient’s response to a treatment usually includes symptoms and signs, physiologic and laboratory measurements, clinical events, side effects, and quality of life.<sup>31</sup> This is true for heart failure trials of drugs and devices.

If multiple primary end points are specified instead of a single outcome or a single composite end point and each

is analyzed separately, the study design and analysis plan must address the increased risk of a type 1 error. A common approach for doing this is to use a Bonferroni procedure and to set the type 1 error level ( $\alpha$ ) such that the probability of incorrectly rejecting at least 1 true null hypothesis is  $\leq \alpha$  irrespective of how many individual null hypotheses are true. With the Bonferroni procedure, if there are  $P$  end points, those with single test  $P$  values  $\leq \alpha/P$  are considered statistically significant.

In some cases it may be desirable to allocate type 1 error unequally across the coprimary end points.<sup>32</sup> For example, EPHEBUS had 2 primary end points.<sup>11</sup> The all-cause mortality end point was tested at the 0.04 level of significance and the composite end point of CVD mortality or hospitalization was tested at the 0.01 level of significance, preserving an overall type 1 error rate of 0.05. With this approach, it is possible that 1 end point would achieve significance and the other would not. In EPHEBUS that was not the case. This could result in the same difficulty of interpretation of study results as composites, particularly if the 2 end points went in different directions.

Whether to “put all of your eggs in 1 basket” can be a difficult decision to make. “Hedging your bet” does not always work. For example, in the Capricorn study, the primary end point was changed from a single outcome, all-cause mortality, to 2 coprimary outcomes, all-cause mortality to be tested at the 0.005 level of significance, and a composite outcome, all-cause mortality or CVD hospitalization, to be tested at the 0.045 level of significance. This change was made because the Data and Safety Monitoring Board noted that the overall mortality was lower than expected and the sample size for that outcome was inadequate based on the design assumptions. After completion of the trial, the  $P$  values corresponding to the all-cause mortality and composite outcomes were .031 and .296, respectively. Neither end point reached prespecified significance levels; however, the all-cause mortality end point would have had the coprimary composite end point not been added. This is also an example of the hospitalization component of the composite end point yielding a much smaller treatment difference (275 versus 289 patients) than the mortality component (116 versus 151 patients).<sup>33,34</sup>

The strength of the Bonferroni procedure (and also less conservative procedures<sup>35</sup>), is also its weakness with respect to multiple outcomes in clinical trials. The Bonferroni procedure has its greatest power in situations when only one of  $P$  end points has a non-zero treatment difference. In most situations it would be considered undesirable if only 1 of several efficacy outcomes used in a clinical trial was significant. If other outcomes did not trend in the same direction, the interpretation of the trial would be complicated. If multiple outcomes are expected to produce consistent results, the Bonferroni procedure is conservative.<sup>36–39</sup> Less conservative procedures for handling multiple end points have been developed. A review of them can be found elsewhere.<sup>40,41</sup>

If the end points are correlated and similar effects are expected across multiple outcomes, a global test such as the one developed by O’Brien can be very useful.<sup>42</sup> O’Brien’s rank-sum procedure is both simple to implement and potentially powerful. With it each of the end points is ranked (eg, greatest benefit to least benefit) noting the treatment group the patient was in. The sum of the ranks across the  $P$  end points for each patient is computed. With 2 treatment groups, the sum of the ranks for each patient can be averaged over the treatment groups and be compared with Student’s  $t$  test or an equivalent statistical test. This procedure has excellent power when all the end points trend in the same direction. A disadvantage is that it is possible to obtain an overall significant difference between treatment groups without any of the individual end points achieving nominal significance. This complicates communication of results. Koch and colleagues have noted that in practice one might have to show significance for at least one of the components as well as for the global test for the results of the study to be accepted.<sup>43</sup>

In a similar vein, Capizzi and colleagues describe a decision rule for an asthma trial with 2 types of outcomes, each with 2 components.<sup>38</sup> One pair of outcomes is based on pulmonary function measurements (forced expiratory volume and peak expiratory flow rate), and the other pair includes self-reported symptoms and self-reported use of asthma medication. They define a rule that requires one end point in each pair to be significant at the 0.05 level and the other to trend in the same direction at the 0.20 or 0.10 level of significance. For moderate correlation among the 4 end points ( $<0.40$ ), the experiment-wise type 1 error rate is substantially less than 0.05.

Tilley and colleagues<sup>44</sup> and Sankoh and colleagues<sup>41</sup> describe the advantages and disadvantages of different global tests. Tilley and colleagues<sup>44</sup> discuss global tests for combining disability scales for stroke trials. Many of the issues considered are relevant to the combination of different functional outcomes for heart failure trials (eg, NYHA class, quality of life and 6-minute walk). Sankoh and colleagues<sup>41</sup> use simulation methods to compare the type I and II error performance of global approaches with other measures for that deal with each outcomes separately.

Follman and colleagues described an approach for handling multiple outcomes whereby the entire event experience during the trial is ranked by expert raters.<sup>45</sup> This is analogous to the idea of ranking the components of the composite that was previously mentioned,<sup>13,24–26</sup> except the entire event profile is ranked. For example, consider the fifth patient in Fig. 2. This patient was hospitalized 3 times and then died. With the approach of Follman and colleagues,<sup>45</sup> this event profile would be ranked along with other profiles (eg, 5 hospitalizations, death without hospitalization) by experts and the average ranks for the treatment groups compared. As noted earlier, different approaches for handling the varying severity of clinical events that patients might experience and using all of the information on events that occur during a trial have been considered by Neaton and

colleagues and Bjorling and colleagues for HIV treatment trials<sup>16,17</sup> and by Chang and colleagues<sup>46</sup> for vaccine trials. Related to this, Metcalfe and colleagues discuss different summary statistics for multiple hospital admissions in heart failure trials.<sup>47</sup> Many of the ideas discussed by these authors are applicable to other disease areas.

### Summary

There are advantages and disadvantages to using a composite outcome versus a single outcome versus using another approach to handling multiple primary outcome measurements. These advantages and disadvantages are summarized in Fig. 3.

A single, clinically relevant, primary end point that is not a composite of several outcomes has the advantage of simplicity. For example, a clear survival benefit for a new heart failure treatment would likely trump any other finding.

Use of a single composite outcome can result in a smaller sample size if the hypothesized effect of the treatment on each component of the composite is similar. If, a priori, there is uncertainty about the consistency of the possible effects of the treatment on the different components, it should not be used as the primary end point. For this reason, it is usually not wise to combine safety and efficacy outcomes into a composite. Such outcomes need to be examined separately to balance risks and benefits.

Alternative approaches to handling multiple outcomes such as a global index and hierarchical scoring or ranking of end points are less used than other approaches to defining composites. In part, this is due to uncertainty about clinical relevance. What may be clinically relevant to 1 group of investigators may not be to another. Where possible, in completed trials, novel composites should be defined and studied before they are used as primary outcomes in new trials. Knowledge of the correlation among the components and the distribution of the novel composite is important to understand before prospective use. Cutter and colleagues<sup>48</sup> describe how data from completed trials of treatments for

multiple sclerosis were used to define a new composite outcome for future multiple sclerosis trials.

In summary, multiple outcomes in clinical trials are a necessity. To understand whether a treatment makes patients feel better and live longer and, if it does, how, clinical, functional, structural, and laboratory outcomes are usually required. In some cases, a composite can be a meaningful and powerful way of combining different outcomes. Even if a composite end point is not the primary outcome, different composite outcomes can be useful as secondary outcomes. Further standardization of composite outcome definitions across heart failure trials would facilitate study comparisons and meta-analyses. When composites are used, all components should be reported. To do this properly, data collection for all components should continue until the end of the trial.

### Acknowledgments

We would acknowledge the following individuals who participated in the panel discussion on composite end points at the Heart Failure Society of America Workshop, "Key issues in end point selection and measurement of devices in heart failure," on April 1–2, 2004: Michael A. Acker, Steve Anderson, Mariell Jessup, Douglas Mann, and Daniel Mark.

### References

1. Meinert CL. Clinical trials dictionary. Baltimore (Md): The Johns Hopkins Center for Clinical Trials; 1996.
2. Friedman LM, Furberg CD, DeMets DL. Fundamentals of clinical trials. Ed. 3. New York: Springer; 1998.
3. Pocock SJ. Clinical trials. A practical approach. New York: John Wiley & Sons; 1997.
4. Meinert CL. Clinical trials. Design, conduct and analysis. New York: Oxford University Press; 1986.
5. Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? *Ann Int Med* 1996;125:605–13.
6. Prentice RL. Surrogate endpoints in clinical trials: definition and operational criteria. *Stat Med* 1989;8:431–40.
7. HIV Surrogate Marker Collaborative Group. Human immunodeficiency virus type 1 RNA level and CD4 count as prognostic markers and surrogate end points: a meta analysis. *AIDS Res Retroviruses* 2000;16:1123–33.
8. Packer M, Carver JR, Rodehoffer RJ, et al. Effect of oral milrinone on mortality in severe chronic heart failure. The PROMISE Study Research Group. *N Engl J Med* 1991;325:1468–75.
9. DeMets DL. The role of surrogate outcome measures in evaluating medical devices. *Surgery* 2000;128:379–85.
10. Yusuf S, Negassa A. Choice of clinical outcomes in randomized trials of heart failure therapies: disease-specific or overall outcomes? *Am Heart J* 2002;143:22–8.
11. Pitt B, Remme W, Zannad F, Neaton J, Martinez F, Roniker B, et al. Eplerenone Post-Acute Myocardial Infarction Heart Failure Efficacy and Survival Study Investigators. Eplerenone, a selective aldosterone blocker, in patients with left ventricular dysfunction after myocardial infarction. *N Engl J Med* 2003;348:1309–21.
12. Yusuf S, Pfeffer MA, Swedberg K, Granger CB, Held P, McMurray JJ, et al. CHARM Investigators and Committees. Effects of candesartan in patients with chronic heart failure and preserved left-ventricular ejection fraction: the CHARM-Preserved Trial. *Lancet* 2003;362:777–81.

**Choosing a Primary Endpoint: Advantages and Disadvantages of Different Choices**

	<u>Advantage</u>	<u>Disadvantage</u>
Single outcome	Simple	Sample size; multiple endpoints are a reality
Single combined endpoint	Sample size	Interpretation not easy if components show different patterns
Co-primary outcomes	Eggs not all in one basket	Sample size and power
Global index	Power	Not easily interpretable
Hierarchical scoring/ranking	Power; clinical relevance	Clinical relevance

**Fig. 3.** Advantages and disadvantages of different primary end point choices.

13. Califf RM, Harrelson-Woodlief L, Topol EJ. Left ventricular ejection fraction may not be useful as an end point of thrombolytic therapy comparative trials. *Circulation* 1990;82:1847–53.
14. Freemantle N, Calvert M, Wood J, Eastaugh J, Griffin C. Composite outcomes in randomized trials. Greater precision but with greater uncertainty? *JAMA* 2003;289:2554–9.
15. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. ICH harmonized tripartite guideline: statistical principles for clinical trials. *Stat Med* 1999;18:1905–42.
16. Neaton JD, Wentworth DN, Rhame F, Hogan C, Deyton L. Considerations in choice of a clinical endpoint for AIDS clinical trials. *Stat Med* 1994;13:2107–25.
17. Bjorling LE, Hodges JS. Rule-based ranking schemes for antiretroviral trials. *Stat Med* 1997;16:1175–91.
18. Montori VM, Permyer-Miralda G, Ferreira-Gonzalez I, Busse JW, Pacheco-Huergo V, Bryant D, et al. Validity of composite end points in clinical trials. *BMJ* 2005;330:594–6.
19. Poole-Wilson PA, Lubsen J, Kirwan BA, van Dalen FJ, Wagener G, Danchin N, et al. A Coronary disease Trial Investigating Outcome with Nifedipine gastrointestinal therapeutic system investigators. Effect of long-acting nifedipine on mortality and cardiovascular morbidity in patients with stable angina requiring treatment (ACTION trial): randomized controlled trial. *Lancet* 2004;364:849–57.
20. Bristow MR, Saxon LA, Boehmer J, Krueger S, Kass DA, De Marco T, et al. Comparison of Medical Therapy, Pacing, and Defibrillation in Heart Failure (COMPANION) Investigators. Cardiac-resynchronization therapy with or without an implantable defibrillator in advanced chronic heart failure. *N Engl J Med* 2004;350:2140–50.
21. Copeland JG, Smith RG, Arabia FA, et al. Cardiac replacement with a total artificial heart as a bridge to transplantation. *N Engl J Med* 2004;351:859–67.
22. The Pimobendan in Congestive Heart Failure (PICO) Investigators. Effect of pimobendan on exercise capacity in patients with heart failure: main results from the Pimobendan in Congestive Heart Failure (PICO) trial. *Heart* 1996;76:223–31.
23. Lubsen J, Kirwan B. Combined endpoints: can we use them? *Stat Med* 2002;21:2959–70.
24. Hallstrom AP, Litwin PE, Weaver WD. A method of assigning scores to the components of a composite outcome: an example from the MITI trial. *Cont Clin Trials* 1992;13:148–55.
25. Topol EJ, Califf RM, Van de Werf F, et al. Perspectives on large-scale cardiovascular clinical trials in the new millennium. *Circulation* 1997;95:1072–82.
26. Packer M. Proposal for a new clinical end point to evaluate the efficacy of drugs and devices in the treatment of chronic heart failure. *J Cardiac Failure* 2001;7:176–82.
27. Leon MB, Teirstein PS, Moses JW, Tripuraneni P, Lansky AJ, Jani S, et al. Localized intracoronary gamma-radiation therapy to inhibit the recurrence of restenosis after stenting. *N Engl J Med* 2001;344:250–6.
28. Sapirstein W, Zuckerman B, Dillard J. FDA approval of coronary artery brachytherapy [editorial]. *N Engl J Med* 2001;344:297–9.
29. Franciosa JA, Taylor AL, Cohn JN, Yancy CW, Ziesche S, Olukotun A, et al. A-HeFT Investigators. African-American Heart Failure Trial (A-HeFT): rationale, design, and methodology. *J Cardiac Failure* 2002;8:128–35.
30. Taylor AL, Ziesche S, Yancy C, Carson P, D'Agostino R Jr, Ferdinand K, et al. African-American Heart Failure Trial Investigators. Combination of isosorbide dinitrate and hydralazine in blacks with heart failure. *N Engl J Med* 2004;351:2049–57.
31. Pocock SJ. Clinical trials with multiple outcomes: a statistical perspective on their design, analysis, and interpretation. *Cont Clin Trials* 1997;18:530–45.
32. Moyé LA. Alpha calculus in clinical trials. Considerations and commentary for the new millennium. *Stat Med* 2000;19:767–79.
33. The Capricorn Investigators. Effect of carvedilol on outcome after myocardial infarction in patients with left-ventricular dysfunction: the Capricorn randomized trial. *Lancet* 2001;357:1385–90.
34. Dargie HJ. The Capricorn Steering Committee. Design and methodology of the CAPRICORN trial—a randomized double blind placebo controlled study of the impact of carvedilol on morbidity and mortality in patients with left ventricular dysfunction after myocardial infarction. *Eur Heart J* 2000;1:325–32.
35. Hochberg YA. Sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 1988;75:800–2.
36. Pocock SJ, Geller NL, Tsiatis AA. The analysis of multiple endpoints in clinical trials. *Biometrics* 1987;43:487–98.
37. Follmann D. Multivariate tests for multiple endpoints in clinical trials. *Stat Med* 1995;14:1163–75.
38. Capizzi T, Zhang J. Testing the hypothesis that matters for multiple primary endpoints. *Drug Inform J* 1996;30:949–56.
39. Schulz KF, Grimes DA. Multiplicity in randomized trials I: endpoints and treatments. *Lancet* 2005;365:1591–5.
40. Zhang J, Quan H, Ng J, Stepanavage ME. Some statistical methods for multiple endpoints in clinical trials. *Cont Clin Trials* 1997;18:204–21.
41. Sankoh AJ, Huque MF, Russell HK, D'Agostino RB. Global two-group multiple endpoint adjustment methods applied to clinical trials. *Drug Inform J* 1999;33:119–40.
42. O'Brien PC. Procedures for comparing samples with multiple endpoints. *Biometrics* 1984;40:1079–87.
43. Koch GG, Gansky SA. Statistical considerations for multiplicity in confirmatory trials. *Drug Inform J* 1996;30:523–34.
44. Tilley BC, Marler J, Geller NL, Lu M, Legler J, Brott T, et al. Use of a global test for multiple outcomes in stroke trials with application to the National Institute of Neurological Disorders and Stroke t-PA stroke trial. *Stroke* 1996;21:2136–42.
45. Follmann D, Wittes J, Cutler JA. The use of subjective rankings in clinical trials with an application to cardiovascular disease. *Stat Med* 1992;11:427–37.
46. Chang MN, Guess HA, Heyse JF. Reduction in burden of illness: a new efficacy measure for prevention trials. *Stat Med* 1994;13:1807–14.
47. Metcalfe C, Thompson SG, Cowie MR, Sharples LD. The use of hospital admission data as a measure of outcome in clinical studies of heart failure. *Eur Heart J* 2003;24:105–12.
48. Cutter GR, Baier ML, Rudick RA, Cookfair DL, Fischer JS, Petkau J, et al. Development of a multiple sclerosis functional composite as a clinical trial outcome measure. *Brain* 1999;122:871–82.