Variable Impact of Combining Fatal and Nonfatal End Points in Heart Failure Trials

Hicham Skali, MD, MSc; Marc A. Pfeffer, MD, PhD; Jacobus Lubsen, MD, PhD; Scott D. Solomon, MD

Abstract—Randomized clinical trials (RCTs) are a cornerstone of evidence-based medicine. Their design, implementation, and interpretation are sometimes subject to flaws and errors. To achieve statistical significance and economically feasible RCTs, the use of composite end points in heart failure (HF) trials has become more common. We analyzed the incremental value of combining HF hospitalizations with all-cause mortality in trials of chronic HF that enrolled >1000 placebo patients and had a mean follow-up >9 months. We tested the assumption that, compared with mortality, combining HF hospitalization with all-cause death would yield a consistently predictable increase in event rate across HF RCTs. Average placebo arm duration of follow-up was determined, and standardized placebo event rates per 100 patient-years were estimated. Twelve major HF RCTs were included in this analysis. There was a substantial relative increase in the event rate ranging from 64% to 134% when a composite end point was used. This increase was not related to disease severity as described by annual mortality rate. The relative contribution of combining HF hospitalization with all-cause mortality was, however, influenced by the duration of the trial. Combining HF hospitalization with all-cause mortality increases the overall event rate in HF clinical trials; the relative increase varies widely and is unrelated to disease severity. Longer-duration trials have a more predictable increase in events than short RCTs. Trial duration must be considered when composite end points are used during the design and interpretation of HF RCTs. (Circulation. 2006;114:2298-2303.)

Key Words: heart failure ■ trials ■ methods

Death from all causes is the most uncontroversial end point for a clinical trial, and a reduction in mortality rate based on a reliable number of events is generally considered the greatest achievement for any therapeutic intervention.1,2 Moreover, there is no subjectivity in the definition and ascertainment of this end point. The Vasodilator Heart Failure Trial (V-HeFT)3 and Cooperative North Scandinavian Enalapril Survival Study (CONSENSUS),4 the first major heart failure (HF) randomized clinical trials (RCTs) demonstrating a survival benefit, enrolled only 642 and 253 patients, respectively. Less than 15 years later, the Valsartan Heart Failure Trial (Val-HeFT)5 and Candesartan in Heart Failure—Assessment of Reduction in Mortality and Morbidity (CHARM),6 2 of the most recent HF RCTs with all-cause mortality as a primary end point, required 5010 and 7599 patients, respectively, to test the efficacy of angiotensin receptor blockade in reducing mortality in chronic HF patients. With the increasing improvements in prognosis with current combinations of therapies,7–9 HF trials using all-cause mortality as a primary end point would either require much larger sample sizes or alternatively would be restricted to only the most severely ill cohorts.10,11

To test the hypothesis that a novel therapy offers clinical value in a stable, modernly managed population, composite end points have been used more frequently as a primary outcome measure for HF RCTs. Combining death with HF hospitalization is attractive because this nonfatal outcome is clinically and economically important and is associated with a subsequent higher risk of death.12,13 Moreover, the nonfatal aspect of this composite end point is meaningful, definable, and a disease-specific target for a potential therapy to improve clinical outcomes in HF.

When the composite of all-cause mortality or HF hospitalization is used, however, an assumption must be made in the design phase of the RCT regarding the incremental number of events that would be added when this composite end point is used instead of all-cause mortality. We conducted an analysis of the incremental benefits achieved by using this composite end point in HF RCTs.

Methods

We reviewed major placebo-controlled trials of chronic HF that randomized at least a thousand placebo patients, had an average follow-up of >9 months, and reported both mortality and hospitalizations for HF. A search of the literature with the use of MEDLINE, cross-referenced articles, and the Cardiovascular Trials Review (6th edition) by Kloner and Birnbaum was performed. The following data were required for this analysis: (1) number of deaths, (2) number of
patients who experienced the composite end point of death or worsening HF, (3) duration of follow-up based on mortality in patient-years, (4) duration of follow-up based on the composite end point in patient-years, and (5) mean duration of the trial. These data were either extracted from published articles or obtained by contacting the investigators for supplemental information.

The duration of follow-up for these RCTs varied considerably; hence, standardized event rates were calculated for comparisons. For a constant (absolute) hazard (expressed as the number of events per unit of patient time “at risk” of event), the event-specific risk depends on the duration of follow-up and follows from \( S(t) = \exp(-h.t) \), where \( h \) is the event hazard, \( S(t) \) is the proportion of patients that is event-free, and \( t \) is the duration of follow-up. Accordingly, reported proportions of patients with events can only be compared when duration of follow-up is similar. We made the additional assumption that the hazard rate in each treatment arm of each trial was constant over time. For a chronic disease and a relatively short follow-up (no effect of aging), this is a reasonable assumption.14

In the present analysis we considered the hazard rate for the following 2 end points: (1) all-cause mortality and (2) all-cause mortality or first hospitalization for HF (combined end point). Only the Metoprolol CR/XL Randomized Intervention Trial in Congestive Heart Failure (MERIT-HF)15 reported directly the necessary data to estimate hazards rates for these end points (duration in patient-years of follow-up for each end point). For 6 other trials, data were provided by the investigators. For all other trials, we calculated hazard rates by treatment arm for the 2 endpoints of interest on the basis of information available in the article. Because there is no uniformity in the manner in which results of trials are reported,16 we used several different methods and sometimes had to make additional assumptions. Full details of our calculations for each trial are given in the Appendix (in the online-only Data Supplement). Briefly, we based our calculations on (1) data obtained from figures and numbers displayed below Kaplan-Meier curves; (2) hazard estimates from Kaplan-Meier curves; and (3) solving 2 equations with 2 unknowns where the 2 unknowns are the arm-specific length of follow-up, with the equations containing the combined mean duration of follow-up, hazard ratio, and number of patients and events per treatment arm.

To avoid the different therapeutic effects being investigated in each of these trials, the analyses were based exclusively on the placebo arm event rates. The contribution of adding HF hospitalization to all-cause mortality was analyzed in 2 ways: (1) the percent increase was defined as the relative difference between the event rate of the composite end point and of all-cause mortality [(composite − mortality)/mortality]; and (2) the increment rate was defined as the difference between the composite end point event rate and the mortality event rate, both per 100 patient-years of follow-up. The placebo mortality rate per 100 patient-years of follow-up was used to characterize the severity of the disease in each of the trials. Duration of the trial, if not specified by the authors, was derived on the basis of the starting and ending dates of patient enrollment and end of trial (half the accrual period added to the follow-up period). Otherwise, it was estimated as the mean follow-up as stated by the authors.

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Mean Trial Duration, mo</th>
<th>No. Randomized to Placebo</th>
<th>Placebo No. of Events (Cumulative Incidence %)</th>
<th>Corresponding Placebo Event Rates (No. of Events/100 Patient-Years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>40.5†</td>
<td>1509</td>
<td>237 (15.7) 276 (18.3) 411 (27.2)</td>
<td>5.4 10.3 4.9 91</td>
</tr>
<tr>
<td>B</td>
<td>36.2*</td>
<td>2117</td>
<td>334 (15.8) 273 (12.9) 518 (24.5)</td>
<td>5.7 9.4 3.7 65</td>
</tr>
<tr>
<td>C</td>
<td>23.0†</td>
<td>2499</td>
<td>484 (19.4) 462 (18.5) 801 (32.1)</td>
<td>9.9 18.6 8.7 88</td>
</tr>
<tr>
<td>D</td>
<td>13.5†</td>
<td>2001</td>
<td>217 (10.8) NA 439 (21.9)</td>
<td>11.0 23.9 12.9 117</td>
</tr>
<tr>
<td>E</td>
<td>44.5†</td>
<td>1272</td>
<td>412 (32.4) 356 (28.0) 587 (46.2)</td>
<td>11.1 18.2 7.1 64</td>
</tr>
<tr>
<td>F</td>
<td>36.3†</td>
<td>1015</td>
<td>296 (29.2) 286 (28.2) 433 (42.7)</td>
<td>11.5 19.4 7.9 69</td>
</tr>
<tr>
<td>G</td>
<td>42.3*</td>
<td>3403</td>
<td>1194 (35.1) 1180 (34.7) 1781 (52.3)</td>
<td>12.0 22.1 10.1 84</td>
</tr>
<tr>
<td>H</td>
<td>13.1*</td>
<td>1320</td>
<td>228 (17.3) 197 (14.9) 359 (27.2)</td>
<td>13.3 27.6 14.3 108</td>
</tr>
<tr>
<td>I</td>
<td>41.9*</td>
<td>1284</td>
<td>510 (39.7) 470 (36.6) 736 (57.3)</td>
<td>14.4 23.6 9.2 64</td>
</tr>
<tr>
<td>J</td>
<td>14.3*</td>
<td>2884</td>
<td>509 (27.6) 614 (21.3) 973 (33.7)</td>
<td>14.3 32.5 17.7 120</td>
</tr>
<tr>
<td>K</td>
<td>11.1*</td>
<td>1133</td>
<td>190 (16.8) 268 (23.7) 357 (31.5)</td>
<td>19.9 46.5 26.6 134</td>
</tr>
<tr>
<td>L</td>
<td>9.4†</td>
<td>1283</td>
<td>242 (18.5) 237 (18.5) 382 (29.8)</td>
<td>26.6 44.6 18.0 68</td>
</tr>
</tbody>
</table>

Letters represent the abbreviations used in the figures. NA indicates not available.

*Estimated mean duration of follow-up for survivors.
†Data supplied by investigators.
‡Complete data available: minimal derivations or assumptions were made.
§Incomplete available data: derivations and assumptions were made.
¶Complete data available: minimal derivations or assumptions were made.

Results

Twelve major randomized, placebo-controlled, double-blind HF trials,5,15,17–26 fulfilled all prespecified criteria. Data were gathered from published articles5,15,17–24 or personal communication with the authors.5,19,21–26 Characteristics and results of selected trials are shown in the Table. The average duration of follow-up of each trial ranged from 9.4 to 44.5 months. Overall, crude cumulative incidence of mortality (unadjusted for duration of observation) in the placebo arm varied from 10.8% in MERIT-HF to 39.7% in Studies of Left Ventricular Dysfunction– Treatment (SOLVD-T). The cumulative incidence of hospitalization for HF ranged from 12.9% in SOLVD-Prevention (SOLVD-P) to 36.6% in SOLVD-T. There was a strong positive correlation between these 2 end points (all-cause mortality and HF hospitalization; r=0.93, P<0.0001; Figure 1). When standardized for duration of observation, mortality in the placebo group ranged from 5.4 in CHARM Preserved to 26.6 in the Vesnarinone Trial.
(VEST; Table), whereas the composite end point of death or HF hospitalization ranged from 9.4 in SOLVD-P to 46.5 events per 100 patient-years in the Carvedilol Prospective Randomized Cumulative Survival (COPERNICUS) trial.

The use of a composite end point yielded a relative increase in the event rate ranging from 64% to 134% (mean [SD], 89% [25%]). The percent increase in the end point rate was not correlated with the severity of the illness as characterized by the annual placebo mortality rate ($r=0.15$, $P=0.64$; Figure 2). Despite a 5-fold difference in mortality rate between SOLVD-P and VEST, there was a similar relative increase in the end point incidence with the use of the composite end point.

In terms of annual incremental event rate, SOLVD-P had the smallest annual increment with 3.7 added events per 100 patient-years, whereas in COPERNICUS there were 26.6 additional events per 100 patient-years of follow-up. When analyzed by length of trial in months, the increment rate was clustered in the range of 3.7 to 10.1 additional events per 100 patient-years in the longer-duration trials (>18 months) compared with a greater variability of the incremental rate in the shorter-duration trials (<18 months), ranging from 12.9 to 26.6 additional events per 100 patient-years (Figure 3).

**Discussion**

RCTs are believed to be the cornerstone of evidence-based medicine. However, a number of errors that may occur during the design and reporting phases of a RCT have the potential to attenuate the conclusions of an experiment. One element of these is sample size estimation and the use of composite end points. An adequate number of subjects need to be recruited and followed so that their end points are sufficient to test the study hypothesis. One of the first steps in designing a clinical trial is estimating this sample size. This is based on the targeted significance level ($\alpha$ or type I error, usually 5%), the desired power (1-$\beta$ or type II error, usually 85% to 95%), the control group event rate, and the magnitude of the expected difference to be detected. With the use of these parameters ($\alpha=5\%$, power=90%, risk reduction=20%), a theoretical increase in the control arm event rate from 10% to 20% would reduce the overall sample size of the trial from 9000 to 4000 patients, which is a significant and attractive objective. Nevertheless, the sample size calculation is an estimation that is based on hypothesized numbers and presumed approximations obtained from prior studies and/or clinical knowledge. The emergence of the composite end point in clinical trials stemmed from, among other requirements, the need to increase the event rate of an individual end point to arrive at a reasonable (and economically feasible) sample size.

This approach has already been used in a variety of cardiovascular RCTs. Studies of fibrinolysis in acute coronary syndromes required a continuing increase in sample sizes to test for an alteration in all-cause mortality as a result of the continuing improvement in short-term treatment strategies. A relatively recent study comparing 4 thrombolytic
strategies in acute myocardial infarction enrolled 41,000 patients to demonstrate a survival advantage in contrast to earlier mortality trials. In the current era, with excellent short-term prognosis, trials of acute coronary syndromes would require extremely large sample sizes to test for an exclusive mortality benefit. Braunwald et al. introduced the notion of the “unsatisfactory outcome” end point for fibrinolysis trials. This approach combines mortality and other “undesirable” outcomes (recurrent myocardial infarction, development of cardiogenic shock or of severe sustained HF, and development of severe bleeding) after the administration of a fibrinolytic therapy. This end point can be used either as a dichotomous end point, assessing whether a patient experienced any of the prespecified events, or as a score conferred to each patient on the basis of the worse event experienced. Both methods allow for an increase in event rates and statistical power. Likewise, in 4 early major lipid-lowering therapy RCTs with statins, only 2 were designed with death as the primary end point outcome measure. Since then and more recently, the vast majority targeted a composite end point that generally includes at least 1 nonfatal and often multiple end points deemed clinically important and reflecting the presumed target of the investigated therapy.

In our analyses of the 12 selected RCTs in HF, only based their sample size calculations on a composite end point, and the others based theirs on the incidence of deaths. Our data suggest that when a composite end point that adds hospitalizations for HF is used, the number of events that can be added greatly increases event rates (90% increase on average). However, the incremental change varies widely and was more influenced by the length of the trial than by the annual mortality rate as a marker of disease severity. RCTs with a longer duration of follow-up (>18 months) were associated with a more consistent incremental number of events ranging from 4 to 10 added events per 100 patient-years, whereas the contribution from the nonfatal component was much more variable in shorter trials. In RCTs with longer follow-up, the competing-risks effect is more likely to be encountered, and the statistical benefit of a composite end point might be masked. This may be explained by the fact that the longer that HF patients are monitored, the greater is the risk of death, especially in those who have had a hospitalization for HF. In a time-to-first-event analysis, the occurrence of subsequent events does not increase the power or affect the results. Moreover, limiting the reporting to data on first hospitalizations may result in loss of clinically meaningful information about recurrent admissions and overall burden of disease.

Although a theoretical reason for combining multiple individual end points into a composite end point is to increase statistical power, clinical significance may be distorted. To be clinically meaningful and interpretable, all components of the composite end point should be affected in a similar direction. The rationale for combining end points is that the individual end points reflect the same pathophysiological properties of the disease and are similar pharmacological targets of the therapy. However, RCTs may show different results or explain new mechanisms of diseases and treatments. In atherosclerotic disease, it has been widely accepted that the same pathophysiological phenomena across major arteries (coronaries, carotid, cerebral) could justify the use of a composite end point combining stroke and myocardial infarction (fatal and nonfatal) in some antiplatelet trials. Furthermore, in HF trials, combining death with a nonfatal event such as HF hospitalization may camouflage possible detrimental or neutral effects with regard to death. In the analysis of the effect of therapy on a composite end point, one should determine whether (1) the effect is uniform across all components of the combined end point, (2) the effect is mostly carried through the most important clinical end point, or (3) the effect is mostly carried through a clinically less important end point. Accordingly, careful analyses of the individual events clustered in a composite end point should be conducted and reported.

The present study is limited by the number of trials included and the assumptions made to derive some of the data, leading to some uncertainty related to the duration of follow-up for each end point. One of these necessary suppositions was that the composite end point of death and/or HF had a constant event rate. This is secondary to the lack of uniformity in clinical trials reporting, particularly when combined end points are used. Space limitations in journals make publication of detailed results difficult. However, with the availability of the World Wide Web, publication of appendices with detailed methods and results should facilitate this process. Another limitation is the variation in the definition of HF hospitalization across trials. Specifically, RCTs may have used different criteria for definition of HF hospitalization or death and/or HF, which may lead to some misclassification when event rates are compared across trials. In addition, not all trials used a central end points adjudication committee, which also could lead to random misclassification. The trials selected in this analysis span more than a decade and reflect therapeutic changes that have affected baseline treatment in the placebo arm throughout this time. Despite this, placebo event rates are similar in SOLVD-T, the Digitalis Investigation Group (DIG) trial, Cardiac Insufficiency Bisoprolol Study II (CIBIS II), CHARM Added, and CHARM Alternative, and, in this analysis, timing of the trial does not seem to be an important confounder.

In conclusion, combining HF hospitalization with all-cause mortality into a composite end point in a HF trial may result in a substantial increase in the event incidence; however, this increase is not uniformly proportional to the severity of the HF. Longer-duration trials yield smaller and more predictable increases. Hence, the sample size required for a short trial may not be estimated reliably on the basis of the composite end point event rate. These factors should be taken into account when clinical trials are designed and when the results of trials using multiple composite end points are reported.

Acknowledgments

The authors would like to thank all the investigators who shared their data about the following trials: CHARM, CIBIS II, COPERNICUS, DIG, and Omapatrilat Versus Enalapril Randomized Trial of Utility in Reducing Events (OVERTURE).
Disclosures
None.

References
37. Design features and baseline characteristics of the LIPID (Long-Term Intervention with Pravastatin in Ischemic Disease) Study: a randomized trial in patients with previous acute myocardial infarction and/or unstable angina pectoris. Am J Cardiol. 1995;76:474–479.


