Review: Clinical Trial Methodology

CLINICAL INVESTIGATION

Rethinking Phase II clinical trial design in heart failure

Clin. Invest. (2013) 3(1), 57–68

Kory J Lavine & Douglas L Mann*

Center for Cardiovascular Research, Division of Cardiology, Department of Medicine, Washington University School of Medicine, Campus Box 8086, 660 South Euclid, St Louis, MO 63110, USA
*Author for correspondence:
Tel.: +1 314 362 8908
E-mail: dmann@dom.wustl.edu

The incidence and economic burden of heart failure continue to rise worldwide, despite implementation of a number of effective heart failure therapies. Although there have been a number Phase I–II studies of potential novel heart failure therapies over the past decade, none of these new compounds has been successful in Phase III clinical trials. While there are likely a number of reasons for this failure, one of the problems that has become increasingly apparent is the inability of Phase II trials to correctly identify novel therapies that will be successful in Phase III clinical trials. The following review will discuss some of the problems inherent with current Phase II heart failure clinical trials, as well as possible ways to rethink Phase II development of new therapies for heart failure.

Keywords: adaptive study design • clinical translation • clinical trial design • heart failure • Phase II

Heart failure is a burgeoning healthcare problem worldwide and is the leading cause of hospital admissions in the industrial world. Despite implementation of effective heart failure therapies and improved clinical outcomes, both the incidence of heart failure and the burden of disease continue to climb. Following the introduction of ACE inhibitors, angiotensin-receptor antagonists, β-adrenergic blocking agents and aldosterone antagonists, there have been very view new effective pharmaceuticals approved for treatment of heart failure. Indeed, with the exception of ivabradine – currently only approved in Europe – there have been no new drugs approved for heart failure since Bidil™ (fixed-dose hydralazine isosorbide), which was approved in 2005 by the US FDA for a very narrow heart failure indication.

Despite a wealth of basic science reports and early preliminary Phase I–II studies describing potential novel therapies, none of these compounds has been successful in Phase III clinical trials. Whether this disconnect between Phase II and III studies signifies a lack of efficacy for the proposed therapeutic strategies, or alternatively a failure to effectively design clinical studies, is a topic of intense debate. The problem of developing new therapies for heart failure reflects the more general problem of developing novel therapies for all classes of disease. Indeed, in 2004 the FDA's Critical Path Initiative identified "*the increasing difficulty and unpredictability of medical product development*" [101]. A 2006 update announced that *"[the FDA's] outreach efforts uncovered a consensus that the two most important areas for improving medical product development are biomarker development and streamlining clinical trials"* [102].

The following review will discuss some of the problems inherent with Phase II clinical trial designs for the development of novel heart failure therapeutics, as well as possible ways to rethink Phase II development of new therapies for heart failure.

Why heart failure drugs fail

Although the discovery of ACE inhibitors, β-blockers and aldosterone antagonists has resulted in dramatic improvement in the care of heart failure patients, it is sobering to realize that these therapies were investigated in Phase III trials

FUTURE SCIENCE part of fsg

largely based on initial observations from small clinical studies, rather than on carefully designed preclinical programs based on novel target identification in the laboratory. Indeed, attempts to develop new heart failure therapies based on the rational identification of drug targets has resulted in disappointing results in Phase III trials. One obvious explanation is that the low hanging fruit may have already been picked, and that demonstrating the benefit of additional add-on therapies on top of conventional triple therapy (ACE inhibitors, β-blocker and aldosterone antagonist) is exceedingly difficult because the annual mortality is now approximately 5–7% for patients with moderate heart failure. Another explanation is that our current approach to treating heart failure has primarily focused on targeting cell-surface receptors or intracellular receptors [1]. While this approach has worked extremely well for antagonizing various components of the adrenergic and the renin–angiotensin–aldosterone systems, this type of 'reductionist' approach has not worked well for antagonizing other systems (e.g., endothelin, adenosine, tumor necrosis factor). As a case in point, ivabradine, which is the most recently approved therapy for heart failure in Europe, blocks the I_f ion current channel, which is highly expressed in the sinoatrial node. Another potential reason for failure to develop new heart failure therapies is our inability to identify effective therapies in Phase II studies, which abound with false-positive results. Equally concerning is the potential for false-negative results that result in cessation of a promising new therapy in Phase II. Last, there is a growing potential for 'false-neutral' results, which refers to therapies that are as yet untested in the clinic because there is not a clear development path forward for these new agents.

Traditional clinical trial design

The initial steps in the clinical development process for a new heart failure therapy are to provide a bridge between the basic science that originally suggested that a disease causing pathway could be targeted therapeutically, and the definitive studies that convince regulatory agencies that the therapy can beneficially influence outcomes in a patient population. This transition from 'bench-to-bedside' is complicated by numerous factors, but it ultimately requires a 'proof-of-concept' confirmation that the therapy performs in humans in its intended manner and a determination of appropriate doses to allow more widespread testing of hypotheses (i.e., dose selection). For newer targets that have not yet been tested in humans, it is often difficult to pick relevant clinical end points that will demonstrate proof of concept. Table 1 summarizes a variety of approaches to these critical steps in the evaluation of new therapies [2].

Clinical development programs of new heart failure therapies have traditionally been divided into three phases, each with distinct objectives and potential issues in trial design. The purpose of the Phase I studies is to assess the safety of a new therapy in humans, with the specific objectives to characterize the metabolism, pharmacokinetics, pharmacodynamics, dose–response, tolerability, and possible dose-limiting side effects of the therapy prior to further investigation. These studies vary greatly in size, but generally range from 20–80 patients. Phase II studies evaluate the dose and effectiveness of the therapy for a specific indication(s) in patients with the condition of interest and determine the common short-term side effects and risks associated with the drug. In pragmatic terms, Phase II studies are critical in defining the dose to be used in the larger Phase III studies and to confirm the proof of concept of the therapy being tested, as well as providing information on the anticipated magnitude of effect to be studied. Due to the limited time and financial resources that are available during the early phases of development, early studies may evaluate a limited number and range of doses and employ surrogate end points for clinical outcomes. Phase III trials are large studies that are designed to convincingly demonstrate the efficacy of the therapy, to provide safety information for a more complete evaluation of its benefit-to-risk characteristics, and to ultimately define how the therapy should be used. These trials often range from several hundred to thousands of patients. Recently, the distinction between Phase II and III trials has become progressively blurred and less useful, and new study designs have been developed that transition rapidly between Phase II and Phase III studies while retaining much of the study architecture [3].

Types of Phase II trials

Phase II clinical trials can be grouped into a single arm study, versus non-randomized trials versus randomized trials that include a placebo arm (Figure 1). Given that the pathophysiology of heart failure is exceedingly complex, and that the responses to a given therapy are highly variable as a result of the heterogeneity of heart failure patients and presence of multiple medical co-morbidities that may mute the effects of an effective new therapeutic agent, a single arm Phase II trial comparing the clinical status of heart failure patients at baseline and after treatment is not likely to be inadequate for identifying compounds that will be successful in Phase III. The next question in Phase II development of a two arm trial is whether randomization is necessary to identify the effectiveness of the novel agent. Randomized trials have many advantages over studies with nonrandomized concurrent or historical

Table 1. Approaches to the design of Phase II trials.

Approach	Method	Advantages	Disadvantages
Intuitive	Provide drug to select group of investigators who observe effects in open-label, unblinded studies and make recommendations	Rapid, cheap and easy to conduct	Investigators cannot reliably discern clinical benefit Purely subjective
Mechanistic	Administer drug in multiple doses and select dose that has optimal biologic effect thought to best reflect mechanism of action of drug	Often quantitative Testable hypothesis Rational	Limited ability to translate animal preclinical research to human studies while maintaining importance of mechanism Biologic effect of drug may be impossible to measure in patients Drug may have multiple effects that may supercede its intended actions Difficult to determine the degree of the biologic effect required to demonstrate efficacy Short-term biologic effects may not be maintained long term No clear relation between mechanistic efficacy and effect on clinical outcomes
Efficacy pilot	Administer drug in multiple doses and select dose that has optimal effect on an arrary of clinical end points	Clinical relevance	Usually no single end point selected, so determines if 'things are generally going in the right direction' Differences are typically small and trends are often conflicting Does not give benefit–risk information
Safety pilot	Administer drug in multiple doses and select dose that has optimal safety profile	Clinical relevance	Assumes that mechanism of action and clinical efficacy is established Difficult to determine the correct dose of a new drug based on its safety profile Unlikely that safety can effectively be assessed in intermediate sized trial (<500 patients) Does not give benefit–risk information

Adapted with permission from [27] © Elsevier (2000).

controls, including the elimination of bias in the assignment of treatments and the achieving balance of the known and unknown baseline covariates that may influence response. However, randomization increases the size (and hence the cost) and duration of the study, which can be problematic in Phase II.

One possibility that has been employed recently in the development of circulatory assist devices for heart failure patients is the use of a historical (i.e., a registry) or contemporary control group that permits a standard protocol to be used for defining inclusion and exclusion criteria. For example, the ADVANCE trial demonstrated that a new circulatory assist device was non-inferior with respect to the primary end point (survival on the original device, transplant or explant for ventricular recovery at 180 days) when compared with a contemporary group of patients with commercially available pump implanted at the same time, who were enrolled in the INTERMACS study [4]. The most significant limitation of this type of trial design is that the comparability of patients in the contemporaneous control group and treatment group cannot be ensured completely.

A popular approach that has been used to deal with baseline imbalances in historical control groups is the propensity score [5]. In observational studies, treatment selection is often influenced by the baseline characteristics of the subjects. Randomized controlled trials eliminate this problem (in theory) by ensuring that the treatment status is not confounded by the baseline characteristics of the treatment group. Propensity score matching is a statistical technique that attempts to estimate the effect of a treatment by accounting for the covariates that predict receiving the treatment, and thus attempts to reduce the bias that is due to confounding variables. Treatment and control outcomes can be compared within strata or adjusted for their propensity score. The limitation of this approach for Phase II heart failure studies is that it assumes that all of the clinically important differences between the treatment and

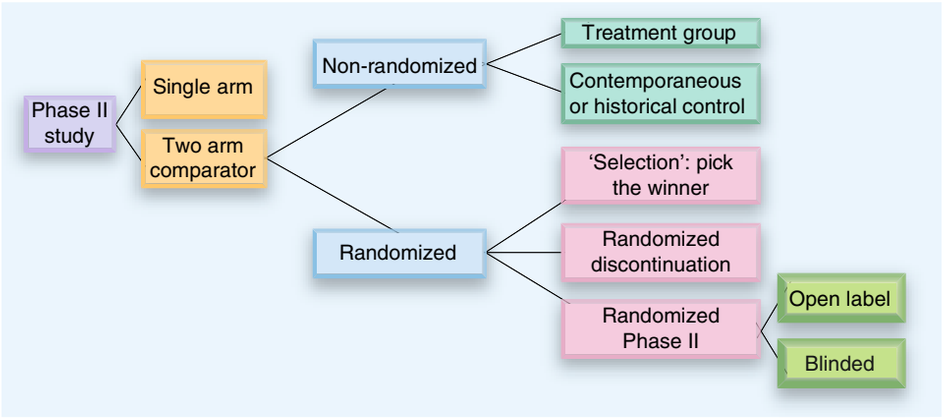

Figure 1. Types of Phase II trial designs.

control groups have been accounted for, and that there are no important but unobserved differences between the two groups.

Based on the aforementioned limitations, randomization is usually the preferred trial design for Phase II trials in heart failure. Given the inherent variability in responses in heart failure, blinding of the agents against placebo, or other doses of the same agent, or other active agents is preferable wherever possible. Unlike many cancer trials, most Phase II heart failure trials proceed directly to either an open label or blinded randomization scheme, rather than employing some of the adaptive steps illustrated in Figure 1. These adaptive approaches will be discussed in detail below.

Phase II trials in heart failure

Typically, Phase II trials in heart failure enroll a modest number of patients (ranging from 50–200), with randomization schemes weighted towards the experimental compound. The inclusion of several study arms leads to a reduction in statistical power and/or leads to a smaller number of patients in the control arm, in order to account for the increased number of enrolled subjects in the treatment arms. However, the collective experience over the past 20 years has shown consistently that there is a price to be paid in heart failure trials when the control group in Phase II or III does not contain a sufficient number of patients. For example, if the event rates in a small control group are spuriously high and not reflective of the event rates for the entire cohort, it may give the false impression that the experimental compound is effective when compared with placebo. Another problem with this approach is that it may lead to a power calculation that overestimates the actual event rates, which can lead to Phase III trial designs that are not sufficiently powered. With either scenario, the likelihood of replicating the results of Phase II in Phase III is substantially less.

There has been considerable debate regarding appropriate clinical end points for Phase II trials. Given that it is not practical to use 'hard' end points such as mortality or mortality plus hospital admissions, most Phase II studies have employed a variety of different surrogates such as functional capacity, quality of life, heart failure symptomatology (measured as the New York Heart Association class), left ventricular (LV) remodeling and/or ejection fraction, biomarkers, or various different clinical composites. A second potential issue in the early phases of development of a novel therapy is there are often limited clinical data available to help guide selection of relevant end points. This often results in 'rounding up the usual suspects', as discussed above, rather than trying to identify target-specific end points. Alternatively, when it is not possible to clearly identify a relevant clinical end point, Phase II trial designs will often employ a panoply of surrogate or composite end points, or changes in biomarkers as secondary end points, in order maximize the chance of finding a positive signal in Phase II. As a result of these problems, many Phase II study designs for novel compounds often do not test relevant hypotheses with respect to how a new agent might positively impact patients with heart failure, and/or are unnecessarily expensive.

Several potential explanations have been proposed to explain why the surrogate end points used in Phase II trials do not predict therapeutic effects on mortality and/or morbidity in Phase III heart failure trials. The explanation that has received the widest acceptance, particularly by regulatory agencies, is that the surrogate does not reliably predict the overall effect on the clinical outcome [6]. For example, heart failure progression could proceed through several causal pathways that are not mediated through the surrogate. Thus, although the intervention may have a significant effect on the surrogate, it may have no effect on the relevant disease-causing pathway. One of the classic examples of this scenario in heart failure trials is the use of exercise capacity to assess changes in functional capacity as a surrogate for disease progression. Although pimobendan and floseqinan improved exercise capacity in heart failure patients, both of these drugs substantially increased patient mortality when studied in Phase III trials [7,8].

Another possibility is that the intervention might also indirectly affect the clinical outcome by unintended, unanticipated, and unrecognized mechanisms of

action that operate independently of heart failure. This is seen in heart failure trials with drugs that improve the quality of life in heart failure patients, yet increase mortality in Phase III studies [7,8]. Thus, surrogates have a relatively poor track record of predicting success in large, randomized, controlled heart failure trials. The notable exception to this statement is the assessment of ventricular remodeling, which has been shown to reliably predict clinical outcomes in clinical trials [9]. Unfortunately, this surrogate end point has not yet gained traction with regulatory agencies, despite overwhelming preclinical and clinical evidence that changes in remodeling are involved in the progression of heart failure.

Heterogeneous patient populations with significant variation in co-morbidities and disease severity, and short follow-up periods are also problematic in Phase II clinical trial designs. Although a new experimental compound may be efficacious with respect to a surrogate end point that lies within the disease-causing pathway, the new compound may ultimately fail in Phase III because the burden of non-cardiac comorbidities often dominate the downstream event rates such as hospitalization and death [10]. While this type of problem can often be accounted for in Phase III trials by focusing the primary end point on cardiovascular hospitalizations and/or cardiovascular death, in many instances the co-morbidity may be inextricably linked to the cardiovascular outcomes (e.g., chronic kidney disease). Indeed this issue may be one of the root causes for our inability to develop drugs for patients with heart failure with preserved ejection fraction.

Differences in experimental methodology between basic & clinical science

A key difference between basic and clinical science methodology is the ability of a basic scientist to continuously define and refine the hypothesis in real time, based on the results of completed experiments. In sharp contrast, clinical study designs employing frequentist statistics require that the hypotheses, drug dose, sample populations, and relevant end points are stated precisely prior to trial initiation, in order to maintain the statistical integrity of the trial. As noted above, prior to the initiation of Phase II studies, there may be little information available with respect to which end points and patient populations an experimental compound may provide the most clinical benefit. In this setting, studies are designed using the generic end points discussed above, and/or investigators attempt to extrapolate end points obtained in previous experimental studies in animal models. Although this type of strategy worked extremely well for the development of ACE inhibitors in the SVAE study [11], this type of approach has not worked well for other targets insofar as small animal models do not precisely recapitulate human disease. Thus, the rigidity of current clinical trial designs often leaves little room for adjustments to be made during the study period as clinical experience with the new agent or device is obtained.

Rethinking Phase II clinical trial design in heart failure

How can Phase II trials be redesigned so that they can fulfill their intended purpose of exploring mechanisms of benefit and predicting the efficacy of novel therapeutics in larger Phase III heart failure trials? As discussed above, our current Phase II trial designs in heart failure are excessively rigid in order to preserve statistical integrity and maximize statistical power. In this regard, numerous new trial designs have been proposed in oncology trials, including randomized selection designs ('pick-the-winner'), adaptive designs [12], randomized discontinuation designs [13], and other randomized designs [3]. Prospectively specified adaptive designs are of particular interest in the context of Phase II heart failure studies of new agents wherein very little is known with respect to the appropriate patient population, or the optimal doses and/or schedules are not clearly known at the trial outset. Adaptive designs in such settings should be efficient and may result in improved precision in terms of predicting success in Phase III. Despite the multiplicity of new designs that have been proposed, their inclusion in heart failure trials has been notably absent.

Adaptive trial design

Adaptive trial designs may be divided into prospective designs, continuously adjusted (*ad hoc*) designs, or retrospective designs [14,15]. Prospective adaptive designs include studies where there is a prespecified protocol to alter parameters such as sample size, follow-up time or clinical end points if a certain threshold is met. For example, based on interim analyses a study may be considered as futile and terminated prematurely. Alternatively, many interim analyses may suggest that a larger sample population or follow-up time may be required to achieve sufficient power for the primary end points. Sequential trial designs represent prospectively prespecified protocols that allow for interim analyses of the data. Continuous or *ad hoc* adaptive trial designs allow investigators the flexibility to alter several study parameters based on previously observed outcomes. For example, participant inclusion criteria, dosing regimens and end points may be modified based on interim

analyses [15]. This approach provides investigators with the opportunity to refine hypothesis based on observed results. Once such changes are made, subsequent conclusions are derived from combining retrospectively and prospectively collected data. While this approach appears attractive, it is has the possibility of introducing significant bias into the study. Specific statistical tools have been designed to minimize the effect of midstudy changes, which are discussed in more detail below. In general, the impact of potential bias is reduced by maximizing the data that are collected prospectively (i.e., following protocol modifications) [15]. This is analogous to the use of discovery and validation cohorts.

Retrospective adaptive designs provide the greatest extent of flexibility for investigators. This trial design allows investigators to change the primary end point or statistical methodology after the trial has been closed, but prior to unblinding of the study. As one would expect, retrospective designs are the most vulnerable to the influence of bias, insofar as none of the data are collected in a prospective fashion following protocol alterations [14].

Given the limitations of bias, most adaptive trials belong to either the prospective or ongoing (*ad hoc*) categories. In either situation, the use of adaptive designs and appropriate statistical methodology are clearly spelled out in the study protocol. Specific adaptations involve modifications of several key aspects of clinical study designs, including refinements of participant inclusion and randomization, sample size, experimental drug dosing protocols and relevant clinical end points. Several aspects of each of these strategies will be discussed below.

■ Adaptive methodology involving study subjects

There are several applications of adaptive trial designs involving study participants that may be well suited to heart failure trials. Given the complexity and heterogeneity of heart failure patients, it is difficult to be certain that the randomization process will always lead to equally balanced treatment and control arms in small Phase II efficacy studies. Adaptive randomization is a prospective strategy that is employed to improve the efficiency of randomization [16]. In this approach, the probability that a subject may be assigned to a particular arm is dependent on the number of patients previously assigned that share particular characteristics. For instance, investigators may use adaptive randomization techniques to increase the probability that each study arm will include the same number of heart failure patients that have the same New York Heart Functional Class, background device use (e.g., biventricular pacemakers) or co-morbidities (diabetes, hypertension and coronary artery disease). This type of trial design technique should decrease heterogeneity, and thus increase the signal-to-noise ratio with respect to analyses of primary and secondary end point. An additional advantage of adaptive randomization is that it may improve the efficiency of subgroup analyses by assuring that subgroups of interest are equally allocated into each study arm.

■ Adaptive methodology to improve statistical power

Adaptive trial designs may also be utilized to improve study power if unanticipated efficacy signals are observed, or if the statistical power for the intended primary end point is insufficient based on event rates in the trial [17]. For example, if an investigational drug is being evaluated for its impact on LV remodeling in a Phase II study and an interim analysis detects that this agent reduces the incidence of heart failure hospitalizations, the study may be modified to better evaluate this preliminary finding. This can be accomplished by enrolling additional subjects and/or extending the duration of follow up employed to adequately power the study. This approach can be also be used to prematurely terminate a study for futility if there is no efficacy signal after the trial has collected sufficient events.

■ Adaptive methodology to enhance the study population

It is also possible to alter the study population during an ongoing trial in an effort to refine the hypothesis that is being tested. For example, if the novel agent under evaluation is designed to improve LV function and reduce heart failure symptoms, and an interim analysis suggests that only patients with ischemic cardiomyopathy receive benefit, then enrollment criteria can be modified to restrict enrollment to patients with ischemic cardiomyopathy, and the sample size of the trial adjusted based on the anticipated power required to demonstrate a statistical difference in the cohort of patients with ischemic heart disease. Using this methodology, it is possible to explore if there is a particular patient population for which a novel therapy may be best suited. This approach may also be used to remove patient populations that either fail to show significant efficacy or are harmed by the tested therapy. An example of this strategy would include terminating enrollment of patients with moderate or severe chronic kidney disease, if an interim analysis demonstrated lack of efficacy or harm in this subgroup. An important caveat of this approach is the possibility of falsely identifying a positive signal for efficacy or harm in a particular patient subgroup. Alternatively,

there is a similar risk of prematurely concluding that a patient population is receiving inadequate benefit at the time of an interim analysis due to insufficient power.

■ Adaptive methodology to compare drug dosing regimens

Phase II efficacy studies are traditionally challenged by the small sample sizes that accompany the multiple study arms required to explore the range of possible doses of a new therapeutic agent that might be appropriate to take forward into a Phase III clinical trial. As a result of these limitations, typically only two or three doses can be effectively evaluated in a single study. The introduction of adaptive study designs has helped streamline assessment of multiple drug dosing strategies. The most common method used is termed 'drop-the-loser' and has been successfully utilized in oncology studies (Figure 2) [18,19], and is analogous to the approach that clinicians take when uptitrating heart failure medications when patients are non-responsive to lower doses of a medication. In this study design, patients are randomly assigned to one of several different dosing arms or a placebo (standard-of-care) arm. Using a common set of end points, experimental arms that do not meet specific efficacy criteria are successively dropped until a single dosing regimen is selected. Adaptive allocation strategies are employed to selectively assign patients into the placebo or better-performing experimental arms. The end point used is typically a composite of efficacy and safety measures. Of note, this study design is not capable of generating traditional dose–response curves. 'Drop-the-loser' designs are well suited for combined Phase II and III studies. Selection of optimal dosing is identified in the Phase II component through the strategies described above. Once a dose is selected, it can then be brought forward into a larger Phase III trial (Figure 2).

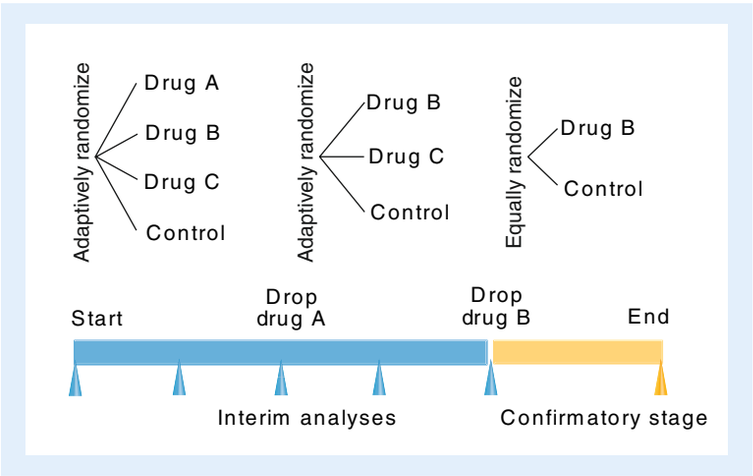

Figure 2. Adaptive trial design for picking the appropriate therapy. In the example trial, single agent drug B is selected in the Phase II part of the trial and continues into Phase III. The number of patients in Phase II is chosen adaptively. The randomization in the Phase II part can also be adaptive, as indicated in the figure. In the Phase III part (confirmatory stage) the sample size depends on the results of Phase II. Phase III might have interim analyses for stopping accrual early, for either expected success or futility. The drug B versus control element during Phase II may (inferentially seamless) or may not (operationally seamless) be counted in the Phase III comparison. Controlling the Type I error rate in the former case requires simulating the entire trial.
Modified and reprinted with permission from [3] © Nature Publishing Group (2011).

■ Adaptive methodology to test multiple hypotheses

Adaptive methodologies can be employed to test multiple hypotheses within a single trial. Analogous to the use of adaptive designs to identify patient subgroups with superior outcomes, adaptive strategies can also be utilized to select appropriate clinical end points from a series of different possible end points. This is a critical design issue, insofar as the ideal clinical efficacy end point may not be immediately obvious at the onset of the trial. Of equal importance, this type of strategy may avoid selecting the wrong primary outcome variable, which may lead to termination of the drug development program. This is an especially important issue for Phase II heart failure trials, as discussed above. Adaptive study designs have the flexibility to prespecify a list of plausible clinical outcome variables in the trial design, and then select the most robust end point through the use of serial interim analysis. Once a primary or small number of primary efficacy end points are identified, the study proceeds in using the prespecified end points defined in the earlier phases of the trial, thus preserving the statistical integrity of the trial. Multiple methodologies may be employed to select from an array of plausible clinical outcomes, ranging from a single interim analysis to multiple interim analyses analogous to the 'drop-the-loser' schemes mentioned previously. The advantage of using these approaches is that investigators do not need to know, or presume to know, the ideal clinical outcome variable that needs to be evaluated. Instead, the selection of primary efficacy end points is driven by observations that are made in the early phases of the trial. It is possible that adaptive designs may not only aid in identifying the most relevant clinical end points to utilize in larger Phase III studies, but also, help define the mechanism(s) by which new therapeutic agents impact on heart failure patients. One disadvantage of this strategy is that a modest set

of plausible clinical outcomes must be prespecified prior to initiation of the clinical study. Including too many outcomes may dilute the statistical power, whereas excluding potential outcomes of interest may limit the information that can be obtained from these types of flexible designs.

■ Statistical considerations for adaptive trial designs

Adaptive trial designs provide an attractive degree of flexibility when compared with traditional frequentist trial designs that employ fixed end points. However, this flexibility comes at the expense of the statistical framework of the trial, which often becomes exceedingly complex in order to preserve the statistical integrity of the trial. Bayesian approaches, which focus on the probability of occurrences, have provided the much-needed technical advances in this field. Details of Bayesian statistics are beyond the scope of this discussion and are reviewed elsewhere [20]. Another method that investigators have used to handle the dilemma of flexible trial design is through the use of interim analysis to evaluate emerging results. This methodological process is referred to as a 'group sequential design'. However, given that repeated analysis of accumulating data with conventional statistical testing can lead to substantially increased false positive (Type I) error rates, it is often necessary to employ an alpha spending function to control the Type I error in order to allow the investigators to evaluate the data as needed. The down side of this is that it requires that a more stringent significance level is used for the primary end point. This obviously poses problems for modest size Phase II trials, which are generally underpowered statistically.

Emerging methodologies for evaluating clinical end points

Composite end points are frequently used in heart failure studies in order to capture more clinical events and thereby increase statistical power. However, there are disadvantages to selecting a composite end point as the primary outcome in a clinical trial. As currently employed, each component of primary end point contributes equally to the composite end point, even though each end point may carry a different clinical significance. For example if the primary end point is a composite of mortality, heart failure hospitalizations, and worsening heart failure symptoms, each of these components is weighted equally. Obviously death is the most important component of the composite. However, reducing mortality may not be achievable in a small Phase II/III trial. Moreover, it is possible that including mortality as an end point in the composite may decrease the statistical power of the primary end point to detect a benefit in the other components of the composite. This was the case in the recently completed Phase III WARCEF trial, which showed a significant (hazard ratio [HR]: 0.5; 95% CI: 0.33–0.82; $p < 0.005$) reduction in ischemic stroke, but was negative for the primary end point of the trial, which was a composite of ischemic stroke, intracranial hemorrhage and all cause death ($p = 0.40$) [21].

How can the process of analyzing composite end points be altered to reflect the relative importance of each component? There have been several distinct approaches proposed to deal with this trial design issue. Two of these strategies are particularly well suited for heart failure trials and are described in detail below.

■ Global rank scoring system

One strategy that has been employed recently is the global rank scoring system, which is based on a hierarchical analysis system, and allows the investigator to appropriately weight components of a composite end point using a global rank scoring system [22]. In the global rank test, study participants are ranked according to a prespecified scheme that weighs clinical outcomes in a hierarchical manner. An advantage of this method is that it can incorporate both clinical events and continuous variables, which allows for the use of biomarkers as a primary outcome variable. For example, consider a trial investigating the effect of a novel therapy on hospitalized patients with decompensated heart failure. The outcomes measured include 30-day mortality, repeat heart failure hospitalization, improvement in heart failure severity, and changes in the level of BNP. At the conclusion of the trial, patients are placed into four groups: subjects who died within 30 days; subjects with repeated admissions for heart failure; patients who were not readmitted with worsening or no change in heart failure symptoms; and patients who were not readmitted with improved heart failure symptoms. Within each group, the patients are ranked by either time to event, or heart failure symptomology and BNP level. Summation of patient ranks is performed for each treatment group to compute the global rank score. The global rank scoring system provides an ordered analysis of the composite end point of 30-day mortality, heart failure readmission, heart failure symptoms and BNP level (Figure 3). A lower score is indicative of worse outcomes as these patients have a higher incidence of mortality, heart failure hospitalizations, worse heart failure symptoms and higher BNP levels. This analysis maximizes the power of using a composite end point without being

subjected to some of the above mentioned pitfalls of traditional approaches.

Win ratio

A second method that can be used to test composite end points in a prioritized manner is the 'win ratio' [23]. This approach is designed to combat two fundamental difficulties that may be present in typical efficacy studies: study population heterogeneity and important events that are censored. The latter is most evident in time-to-event analyses of composite end points, where only the first event is captured and subsequent events are censored. If the first event is less important than the second event, significant information may be lost using current analysis schemes. This is apparent in the above example where the first event was hospitalization and the second event was death. The basis of this methodology is that it forms 'pairs' of patients from the treatment and control arms who are matched on prespecified comorbidities, characteristics of disease severity and duration of follow up, analogous to the technique of propensity matching.

The outcomes of each matched pair are then compared for multiple end points that are ordered in a hierarchical fashion. For example, a representative order might be mortality, heart failure hospitalization, and heart failure symptoms. Using a pre-defined algorithm, matched subjects from the control and treatment arms are designated as a winner or loser (Figure 4) for each of the pre-specified end points. If no winner can be determined, then the pair is considered to be tied. For example, consider a matched pair where subject A died at 3 months and subject B died at 9 months. In this scenario subject B would be the winner. Alternatively, consider another matched pair where subject A was hospitalized at 2 months, and subject B died at 8 months. Here, even though subject A experienced the first event, death is considered a more important outcome and subject A would be the winner. If neither subject A and nor subject B died or were hospitalized, the winner would be determined by heart failure symptom index score. In this example, subjects are considered to be tied if neither subject dies or experiences heart failure hospitalization, and their heart failure symptom index score is identical. Alternatively, a tie would occur if subject A was hospitalized at 9 months and subject B was lost to follow up at 4 months. In this scenario, it is not possible to know if subject B would experience an event (death or hospitalization) prior to subject A.

The total number of wins and losses for each study arm is summed and a win ratio is calculated. The win ratio is defined as the number of wins divided by the number of losses. A p-value and 95% CI can be generated for this type of analysis. This innovative approach was recently proposed and applied retrospectively to several completed randomized, controlled, heart failure trials, including the EMPHASIS-HF [24], and CHARM studies [25]. When the win ratio was applied to each of these clinical trials, the reported results were successfully reproduced using both the matched and unmatched analysis. Although the global rank test and the win ratio have tremendous appeal in terms of Phase II clinical trial design, in that they allow one to evaluate hard clinical end points in Phase II studies, it bears emphasis that these approaches have not been validated in terms of the ability to predict success in Phase III.

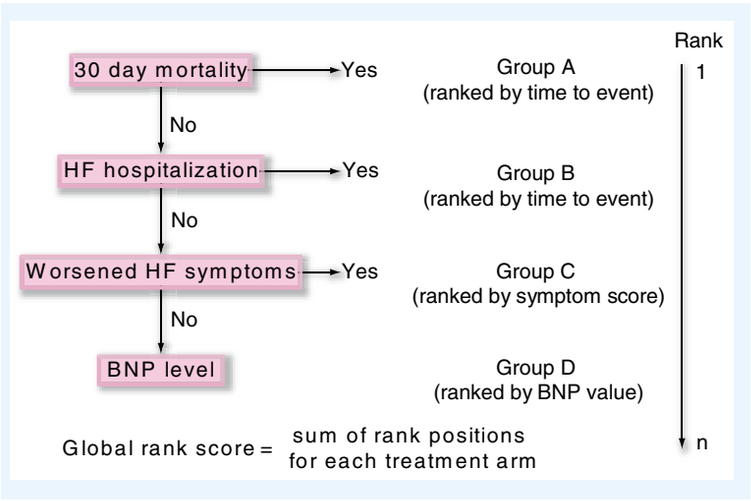

Figure 3. The global rank score. In this example, patients are randomized to receive either a control or experimental therapy. Using the global rank score [22], a composite outcome of 30-day mortality, HF hospitalization, HF symptoms and BNP level is employed. Patients are sorted into four groups using a hierarchical approach: **(A)** experienced mortality within 30 days; **(B)** alive but hospitalized for HF; **(C)** no hospitalizations but experienced worsened HF symptoms; and **(D)** patients without hospitalizations or worsening HF symptoms. Patients in the first two groups **(A–B)** are ranked by time to event. The latter two groups **(C–D)** are ranked by HF severity score and BNP level. A global rank score is obtained for each treatment group by adding the respective patients ranks. A higher score indicates improved outcomes.
HF: Heart failure.

Future perspective

The significant time and cost that must be invested to develop a novel therapeutic using current clinical study designs has become overwhelming from a fiscal perspective, and threatens to dampen enthusiasm for developing novel therapeutics for heart failure, as well as foster the continued development of 'me too' heart failure drugs that target the same pathway in different

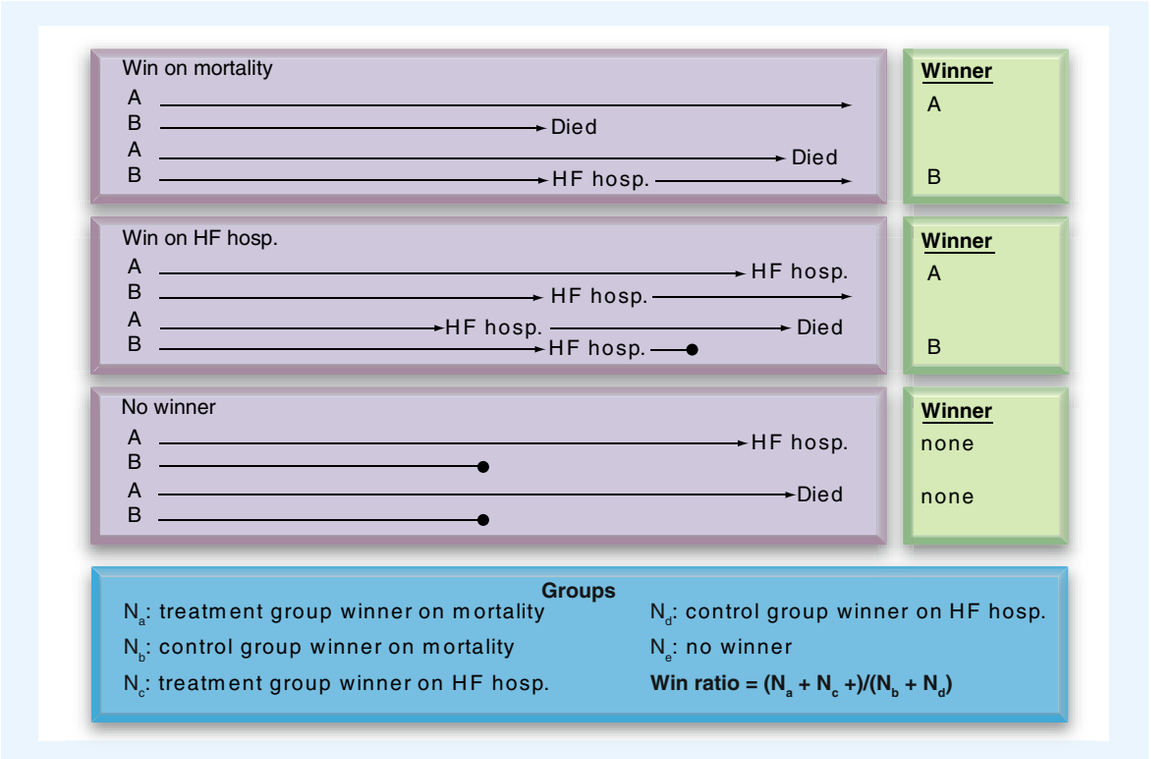

Figure 4. The win ratio. The win ratio is best performed using propensity-matched patients pairs. For each pair a winner and loser is determined, or the pair is considered to be tied. The determination of a winner is made using a predefined hierarchical outcome scheme. In this example, mortality is considered the most important outcome followed by HF hospitalization. Possible scenarios resulting in a winner for mortality and HF hospitalization as well as no winner (or tied) are outlined. The length of each arrow indicates the duration of patient follow up. Arrows ending in a solid circle denote either incomplete or a shorter duration of follow up. The win ratio is calculated by adding the number of wins divided by the number of losses for the experimental group. A statistically significant win ratio that is greater than 1 indicates a positive outcome.
Modified and reprinted with permission from: Pocock SJ *et al*. The win ratio: a new approach to the analysis of composite end points in clinical trials based on clinical priorities. *Eur. Heart J.* 33, 176–182 (2012).
HF: Heart failure; Hosp: Hospitalization.

ways. Phase II trials represent the critical 'portal' that new compounds must pass through prior to entering more definitive Phase III trials. While we agree that the traditional principals of drug development should not be abandoned in Phase III clinical trials, where there is sufficient power to detect differences between placebo and treatment groups for clinically meaningful end points – as noted in the current review – our recent experience has taught us that Phase II heart failure studies in their current form do not predict future success in large Phase III trials. While one could argue that the basic principals of drug development should not be abandoned in Phase II clinical trials, here we argue that Phase II trial designs might be altered to allow for more accurate prediction of lead candidates to move forward to Phase III trials. This opinion is not unique to the authors, but has been championed by the US FDA. To this end, we have reviewed multiple different adaptive approaches that could be employed to improve the efficiency of Phase II studies for indentifying novel therapies for developing heart failure drugs. Although the adaptive trial design approach has been embraced in early-phase oncology trials [3,26], adaptive trial designs have had less uptake by the heart failure community.

While the exact path forward is not at all clear at the time of this writing, there is hope that adaptive trial designs may provide clinical investigators with the flexibility to evolve hypotheses and dosing regimens for novel therapies as they emerge from the laboratory and undergo clinical testing in humans. If clinical investigators are afforded similar luxuries as basic scientists to refine ongoing studies based on prior observations, it is possible that the efficiency and predictive value of Phase II studies may improve remarkably.

Executive summary

Introduction

- The incidence and economic burden of heart failure continue to rise worldwide, despite implementation of a number of effective heart failure therapies.
- There has been a dearth of approved new therapies for heart failure over the past decade.

Why heart failure drugs fail

- Current Phase II heart failure trial designs do not correctly identify novel therapies that will be successful in Phase III clinical trials, we explore:
  - Traditional clinical trial design
  - Types of Phase II trials
  - Phase II trials in heart failure

Rethinking Phase II trial design in heart failure

- Adaptive trial designs offer the promise of improved efficiency and predictive accuracy for developing novel heart failure therapies that are successful in Phase III clinical trials.

Financial & competing interests disclosure

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

No writing assistance was utilized in the production of this manuscript.

References

1 Shah AM, Mann DL. In search of new therapeutic targets and strategies for heart failure: recent advances in basic science. *Lancet* 378, 704–712 (2011).

2 Packer M. Current perspectives on the design of Phase II trials of new drugs for the treatment of heart failure. *Am. Heart J.* 139, S202–S206 (2000).

3 Berry DA. Adaptive clinical trials in oncology. *Nat. Rev. Clin. Oncol.* 9, 199–207 (2012).

4 Aaronson KD, Slaughter MS, Miller LW *et al.* Use of an intrapericardial, continuous flow, centrifugal pump in patients awaiting heart transplantation. *Circulation* 125(25), 3191–3200 (2012).

5 Rubin DB. Estimating causal effects from large data sets using propensity scores. *Ann. Intern. Med.* 127, 757–763 (1997).

6 Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? *Ann. Intern. Med.* 125, 605–613 (1996).

7 Packer M, Narahara KA, Elkayam U *et al.* Double-blind, placebo-controlled study of the efficacy of flosequinan in patients with chronic heart failure. Principal Investigators of the REFLECT Study. *J. Am. Coll. Cardiol.* 22, 65–72 (1993).

8 Packer M, Rouleau J, Swedberg K, Pitt B, Fisher L, Klepper M, and the PROFILE investigators and coordinators. Effect of flosequinan on survival in chronic heart failure: preliminary results of the PROFILE study. *Circulation* 88(Suppl. 1), 301 (1993).

9 Kramer DG, Trikalinos TA, Kent DM. Quantitative evaluation of drug or device effects on ventricular remodeling as predictors of therapeutic effects on mortality in patients with heart failure and reduced ejection fraction: a meta-analytic approach. *J. Am. Coll. Cardiol.* 56, 392–406 (2010).

10 Ather S, Chan W, Bozkurt B *et al.* Impact of noncardiac comorbidities on morbidity and mortality in a predominantly male population with heart failure and preserved versus reduced ejection fraction. *J. Am. Coll. Cardiol.* 59, 998–1005 (2012).

11 Pfeffer MA, Braunwald E, Moye L *et al.* Effect of captopril on mortality and morbidity in patients with left ventricular dysfunction after myocardial infarction. *N. Engl. J. Med.* 327, 669–677 (1993).

12 DeMets DL, Califf RM. A historical perspective on clinical trials innovation and leadership: where have the academics gone? *JAMA* 305, 713–714 (2011).

13 Biswas S, Liu DD, Lee JJ, Berry DA. Bayesian clinical trials at the University of Texas M. D. Anderson Cancer Center. *Clin. Trials.* 6, 205–216 (2009).

14 Chow SC, Chang M. Adaptive design methods in clinical trials – a review. *Orphanet. J. Rare. Dis.* 3, 11 (2008).

15 Chow SC, Chang M, Pong A. Statistical consideration of adaptive methods in clinical development. *J. Biopharm. Stat.* 15, 575–591 (2005).

16 Coad DS, Ivanova A. The use of the triangular test with response-adaptive treatment allocation. *Stat. Med.* 24, 1483–1493 (2005).

17 Hommel G, Bernhard G. A rapid algorithm and a computer program for multiple test procedures using logical structures of hypotheses. *Comput. Methods Programs Biomed.* 43, 213–216 (1994).

18 Ellenberg SS. Select-drop designs in clinical trials. *Am. Heart J.* 139, S158–S160 (2000).

19 Sampson AR, Sill MW. Drop-the-losers design: normal case. *Biom. J.* 47, 257–268 (2005).

20 Berry DA. Bayesian clinical trials. *Nat. Rev. Drug Discov.* 5, 27–36 (2006).

21 Homma S, Thompson JL, Pullicino PM *et al.* Warfarin and aspirin in patients with heart failure and sinus rhythm. *N. Engl. J. Med.* 366, 1859–1869 (2012).

22 Felker GM, Maisel AS. A global rank end point for clinical trials in acute heart failure. *Circ. Heart Fail.* 3, 643–646 (2010).

23 Pocock SJ, Ariti CA, Collier TJ, Wang D. The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. *Eur. Heart J.* 33, 176–182 (2012).

24 Swedberg K, Zannad F, McMurray JJ *et al.* Eplerenone and atrial fibrillation in mild systolic heart failure: results from the EMPHASIS-HF (Eplerenone in Mild Patients Hospitalization And SurvIval Study in Heart Failure) study. *J. Am. Coll. Cardiol.* 59, 1598–1603 (2012).

25 Granger BB, Swedberg K, Ekman I *et al.* Adherence to candesartan and placebo and outcomes in chronic heart failure in the CHARM programme: double-blind, randomised, controlled clinical trial. *Lancet* 366, 2005–2011 (2005).

26 Seymour L, Ivy SP, Sargent D *et al.* The design of Phase II clinical trials testing cancer therapeutics: consensus recommendations from the clinical trial design task force of the national cancer institute investigational drug steering committee. *Clin. Cancer Res.* 16, 1764–1769 (2010).

■ Websites

101 US FDA. FDA's Critical Path Initiative. www.fda.gov/ScienceResearch/SpecialTopics/CriticalPathInitiative/ucmO6689.htm (Accessed 2012)

102 US FDA. Critical Path Opportunities Report. www.fda.gov/downloads/ScienceResearch/SpecialTopics/CrticalPathInitiative/CrticalPathOpportunitiesReports/ucm077254.pdf (Accessed 2012)