"Responder Analysis" for Assessing Effectiveness of Heart Failure Therapies Based on Measures of Exercise Tolerance

DANIEL BURKHOFF, MD, PhD,[1,2] MICHAEL PARIDES, PhD,[1] MARTIN BORGGREFE, MD,[3] WILLIAM T. ABRAHAM, MD,[4] AND ALAN KADISH, MD[5]

New York, New York; Orangeburg, New York; Mannheim, Germany; Columbus, Ohio; Chicago, Illinois

ABSTRACT

Background: Although many studies of heart failure therapies test improvements of patient condition in terms of mean changes of quality of life (QoL) or exercise tolerance (ET) measures, it is of increasing interest to quantify the proportion of patients that "respond" to therapy and understand factors predicting response. These questions can be address through the use of a "responder analysis," in which the proportion of patients in whom a measure of QoL or ET improves by a minimum amount is determined. Here, we review the principles of a "responder analysis."

Methods and Results: We used data from published studies of cardiac resynchronization therapy to model the results of a responder analysis and original data from a recent study of cardiac contractility modulation to illustrate the many facets of such an analysis that need to be understood and investigated further. Some of these areas include: understanding how to choose criteria for response; how to deal with differing results obtained with different measures of response; and how to deal with potentially conflicting information provided by a responder analysis and the more standard comparison of mean changes.

Conclusions: Additional prospective studies will help advance understanding the optimal way to use responder analyses in heart failure trials. (*J Cardiac Fail 2009;15:108–115*)

Key Words: Quality of life, exercise tolerance, cardiac resynchronization therapy, cardiac contractility modulation.

The primary focus of most recent studies of drug therapies for chronic heart failure has been their impact on mortality.[1] However, especially for studies of new therapeutic devices, various measures of quality of life (QoL) and exercise tolerance (ET) have emerged as clinically relevant primary end points. In such studies, overall effectiveness has typically been tested by assessing the difference in mean change over time of a QoL or ET measure between a treatment and a control group. However, it is becoming of increasing interest to understand and quantify the proportion of patients that actually benefit from a therapy. Such a question can be addressed using a "responder analysis," in which the percent of patients in whom a measure of QoL or ET improves by a minimum amount is determined. However, there is relatively little experience with responder analyses as they would apply to these types of outcome measures. Also, it may seem counterproductive and limiting to transform the information contained in continuous variables, which can discriminate varying degrees of response, into a binary "response" or "nonresponse" outcome measure.

Cardiac resynchronization therapy (CRT) is arguably the most important approved device-based therapy for heart failure developed in the past decade.[2–4] Although multiple studies have now shown that CRT improves survival and reduces heart failure hospitalizations, the original US Food and Drug Administration approval of CRT was based on data from the Multicenter InSync Randomized Clinical Evaluation (MIRACLE), which demonstrated significant improvement in New York Heart Association (NYHA) Class, QoL, and ET.[2] In MIRACLE, a traditional analysis evaluating between-group differences in mean changes of the QoL and ET measurements was performed. With respect to NYHA, however, it was reported that 68% of

From the [1]Columbia University, New York, New York; [2]Impulse Dynamics, Orangeburg, New York; [3]Medizinische Fakultät Mannheim der Universität Heidelberg, Mannheim, Germany; [4]Ohio State University, Columbus, Ohio and [5]Northwestern University, Chicago, Illinois.

Manuscript received April 29, 2008; revised manuscript received September 1, 2008; revised manuscript accepted October 14, 2008.

Reprint requests: Daniel Burkhoff, MD, PhD, Division of Cardiology, Columbia University, 177 Fort Washington Ave, New York City, NY 10032. Tel: 201-906-1687 E-mail: db59@columbia.edu

D.B. is an employee of IMPULSE Dynamics, the sponsor of the Fix Heart Failure-4 study of cardiac contractility modulation. M.P., W.T.A., and A.K. are consultants to Impulse Dynamics. M.B. participates in a speakers bureau for IMPULSE Dynamics, Medtronics, and St Jude Medical.

1071-9164/$ - see front matter

doi:10.1016/j.cardfail.2008.10.019

patients in the treatment group exhibited at least a 1 Class improvement in NYHA ranking and that 32% did not show such an improvement. This has led to the often quoted claim that ~70% of patients "respond" to CRT and that approximately 30% of patients are "nonresponders." However, 38% of patients in the control group also exhibited a 1 or more Class improvement in NYHA. Some investigators have therefore argued that the true rate of response to CRT should be corrected to account for this sizeable placebo response in the control group (ie, true response rate = 70 %− 38% = 32%).

To complicate matters, a recent survey of the literature has noted that the rate of response to CRT depends on the variable used to judge response and the criterion used to define response.[5] Changes in 6-minute hall walk distance, peak oxygen consumption (VO_2), QoL (generally using the Minnesota Living With Heart Failure Questionnaire), ejection fraction, end-systolic volume, and the degree of mechanical resynchronization have each been examined and conclusions about response rates vary considerably. In addition, the prominent placebo effect in any heart failure trial creates challenges for any analysis of response.

In this report, we review the principles of a responder analysis as it would be applied to continuous variable outcome measures of ET or QoL. We then use data from the literature to simulate results of a responder analysis for prior studies of CRT. Finally, we use data from a recent study of cardiac contractility modulation (CCM)[6] to test the validity of the underlying principles when applied to an original set of clinical data. The insights derived from these analyses reveal potential benefits and shortcomings of a "responder analysis" and highlight important areas for future research.

Principles of a "Responder Analysis"

To illustrate the principles of a responder analysis, consider the results of a hypothetical clinical trial depicted in Fig. 1. Patients have been randomized between a control (placebo) group and an active treatment group. A variable (such as would quantify ET or QoL) has been measured at baseline (T_o) and at the end of the study (T_{end}). Let us designate the variable of interest as Z, the change in Z for a given patient as ΔZ (ie, $\Delta Z = Z(T_{end}) - Z(T_o)$), the mean change of Z in the population as $\Delta\bar{Z}$ and the differences in means as Δμ ($\Delta\mu = \Delta\bar{Z}_{treatment} - \Delta\bar{Z}_{control}$). In this example, we assume that ΔZ is normally distributed (ie, conforms to a Gaussian distribution) in both groups with common standard deviations. With an effective treatment, the distributions of ΔZ of the 2 groups are shifted relative to each other and a test (such as a *t*-test) is performed to assess whether the shift is statistically significant. Such studies are typically powered (ie, designed to enroll enough patients) to detect the smallest difference in $\Delta\bar{Z}$ between groups that is clinically meaningful.

In the case of a responder analysis, we begin with the same distributions of ΔZ for the 2 study groups (Fig. 1B). A threshold value is set that defines treatment success for each individual patient. The value of this threshold is selected to be equal to or greater than what would be considered to be a clinically meaningful improvement for a patient; selection of the threshold value is a critical matter that, as discussed in detail later in this article, can significantly impact the results. Each patient whose ΔZ exceeds the threshold value is considered a treatment success. Thus the total proportion of patients considered treatment successes in each group equals the area under the distribution to the right of the threshold value (assuming a higher value of ΔZ equates with a better outcome). Then, a statistical test (eg, a chi-squared test) is performed to test whether the proportion of patients meeting the success criterion differs between treatment groups.

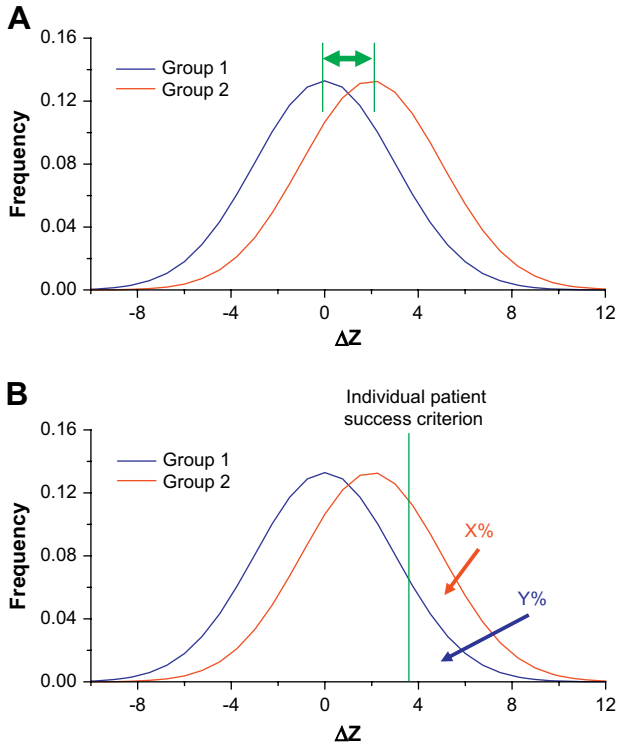

Fig. 1. Hypothetical distributions of changes in a variable Z in a control group (Group 1) and a treatment group (Group 2). In heart failure studies, the analysis typically compares differences in changes mean values (A). However, there is increased interest in understanding the results of a "responder analysis" (B) in which the proportion of patients in which a measure of quality of life (QoL) or exercise tolerance (ET) increases by at least a certain amount is determined.

The information provided is slightly different between these 2 types of analyses. In the case of the comparison of mean values, one can tell the patient the average change in the variable measured he or she should expect to experience if undergoing the treatment. In the case of a responder analysis, one can tell the patient his or her chance of experiencing an improvement in the variable that is equal to or greater than the selected threshold value. The information provided by these 2 types of analyses may or may not be congruent, depending on the nature of the ΔZ distributions.

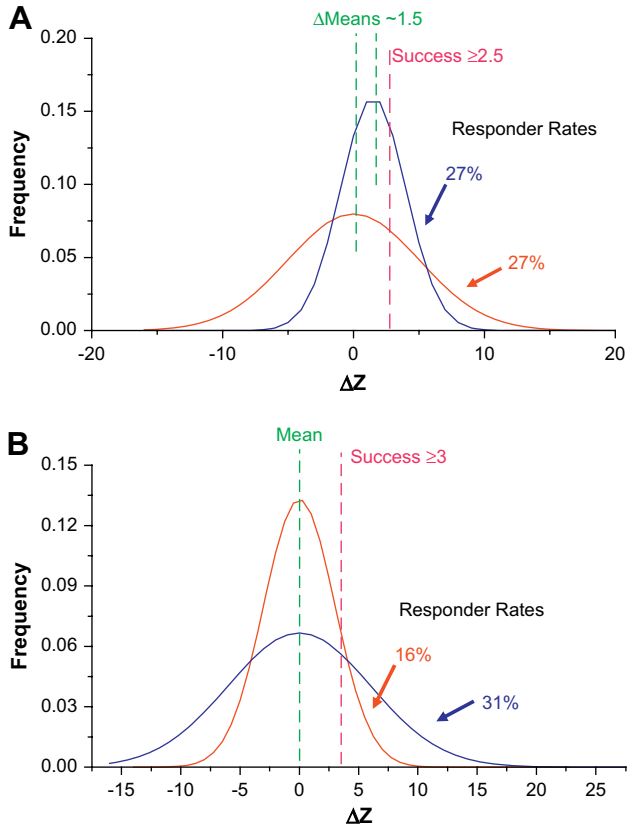

Fig. 2. The results of a "responder analysis" and an analysis of comparison of means could differ. Depending on the nature of the distributions of response, mean values may differ, but the proportion of responders may be the same (A) or mean values may be the same and the proportion of responders may be different (B).

As shown in Fig 2A, the distributions may differ such that the changes in the mean values differ, but the responder analysis indicates a similar response rate. Conversely, as shown in Fig. 2B, the distributions could differ such that there is no difference in the changes in mean values, but there is a substantial difference in responder rates that can be clinically meaningful.

In addition to the considerations discussed previously, it is also important to understand how sample size calculations are influenced by using of a responder analysis compared with an analysis of the changes in mean values. A detailed handling of this question is beyond the scope of this brief report. However, the fundamental principles are revealed through the example illustrated in Fig. 3. We make the following assumptions (Fig. 3A): 1) a randomized study (treatment vs. placebo) tests the effect on an outcome measure "Z", 2) ΔZ is distributed normally in both groups, 3) the mean value of ΔZ in the placebo control group is 0, and 4) the standard deviation of the ΔZ distribution is the same for both groups (in this example, 3 units). Assuming a higher value ΔZ indicates a better outcome, the proportion of patients meeting the success criterion varies inversely with the threshold set to define success, as illustrated in Fig. 3B. As implied by the green arrows, the difference between the response rates between the 2 groups varies directly with the difference in $\Delta\bar{Z}$ between groups. Given these assumptions, Fig. 3C (black squares) shows the sample sizes needed to achieve 90% power as a function of the true difference, $\Delta\mu$, between groups. Sample size decreases monotonically as the difference in the mean change of the distributions increases. For example, if the expected difference between the means is 1 unit, a total sample size of 400 patients (200 per group) is required. The other lines in the graph show the required sample sizes for a responder analysis for different threshold values used to define success criterion as a function of the difference in the means of the distributions. For example, if the expected difference in the means of the distributions is 2 units and the threshold for defining a patient as a "responder" is $\Delta Z \geq 1$ unit, the required sample size is approximately 700 patients (350 per group). As seen in Fig. 3C, in some cases the comparison of means approach requires fewer patients (ie, is a statistically more "efficient" approach) than a responder analysis, but in some cases the responder analysis requires fewer patients.

The general findings summarized in Fig. 3B hold for the case when the distributions are normal, the standard deviations of the distributions of ΔZ are the same in both groups and the mean change of the distribution in the control group is zero. In other cases, the results will be quantitatively different. However, the basic principle remains that in some cases the responder analysis is more efficient, but in other cases the comparison of means is more efficient.

Illustration of Responder Analysis Applied to Studies of CRT

No systematic evaluation of responder rates or comparison of results of the original studies of CRT has been performed previously. Such an evaluation may help to illustrate the various concepts discussed above. One objective measure of ET that has been measured in several studies of CRT is peak VO_2. Two studies in particular readily lend themselves to estimate the results of a responder analysis with this end point: MIRACLE and RHYTHM implantable cardioverter defibrillator (ICD) study.[2,6] In the MIRACLE trial, patients with NYHA III or IV symptoms and prolonged QRS duration were implanted with a CRT device which was turned on in half of the patients and left off in the other half for 6 months in a double-blind manner. The mean ($\pm$ SD) change in peak VO_2 between baseline and 6 months was 0.2 $\pm$ 3.8 mL·kg·min in the control group and 1.1 $\pm$ 3.5 mL·kg·min in the treatment group, so the difference in mean values between the groups was $\sim$0.9 mL·kg·min. Assuming that the data are normally distributed, these numbers can be used to construct estimated distributions of the data (Fig. 4A). In turn, responder rates can be determined for any desired success criterion used to define a responder. It is important to note that there is no objectively justifiable precedence for how to appropriately select a signal value to define success for any ET or QoL measure currently used in heart failure trials. Accordingly, we constructed a graph of the response rate in each group over a broad range of values of success criteria (Fig. 4B). There are 4 important

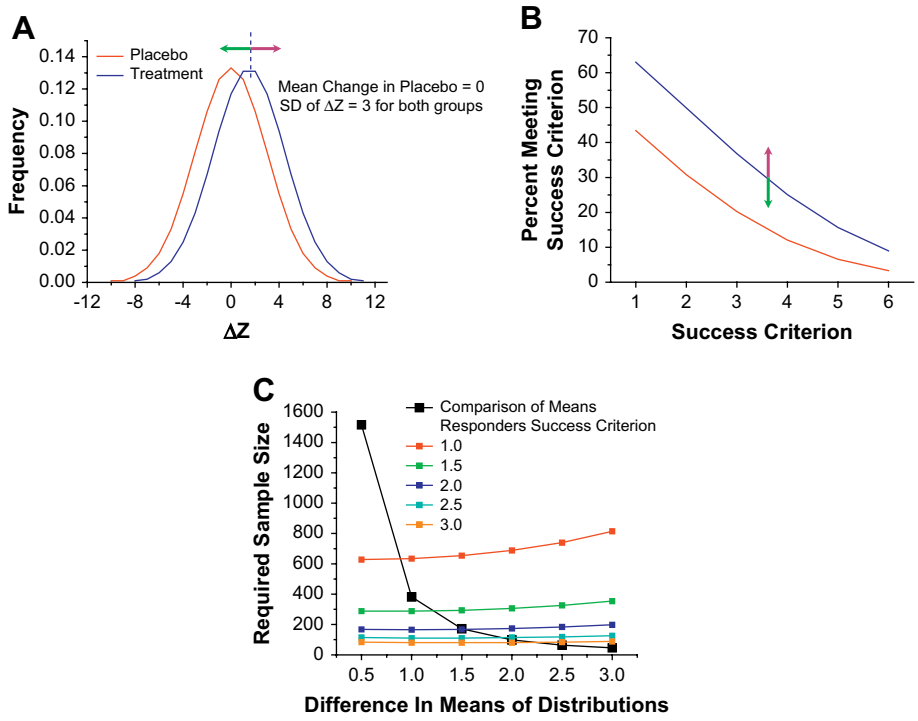

Fig. 3. (A) Hypothetical distributions of changes in a variable "Z" in a Placebo and Treatment group. For this example, standard deviations of the 2 distributions are assumed to be equal. (B) The proportion of "responders" as a function of the success criterion. As the difference between mean values of the 2 groups increases, the difference between the response rates also increases. (C) Comparison of sample size calculations for a "responder analysis" and for a comparison of means analysis. Estimates of required sample sizes are shown as a function of difference in mean values between groups and under other assumptions detailed in the text.

observations. First, there is a clear separation between the control and treatment groups, with a greater number of responders in the treatment group at any value used to define success. Second, the number of responders decreases in both groups as the value of the success criterion increases. Third, the difference in response rate between the treatment and control groups (Fig. 4C) is approximately 10%, and this decreases only slightly with increasing values of the success criterion over the range from 1 to 2.5 mL $O_2 \cdot kg \cdot min$. Fourth, with a success criterion of 1 $mL \cdot kg \cdot min$ (what some might consider a lower boundary for a clinically relevant improvement in peak VO_2) the response rate in the treatment group is approximately 52% and approximately 42% in the control group; the difference between these 2 (ie, 10%) would be the estimate of the net rate of response to therapy.

This analysis can be repeated with data from the multicenter, double-blind RHYTHM ICD study.[7] The mean ($\pm$SD) change in peak VO_2 between baseline and 6 months was -1.4 ± 4.6 mL $O_2 \cdot kg \cdot min$ in the control group and 0.5 ± 2.5 mL $O_2 \cdot kg \cdot min$ in the treatment group, so the difference between the groups was $\sim$1.9 mL $O_2 \cdot kg \cdot min$, a value considerably larger than in the MIRACLE study. Figures 4D–F show the estimated distributions of the change in peak VO_2 for each group, the rate of response as a function of the success criterion and the difference in responder rates between the 2 groups as a function of success criterion, respectively. As can be seen, the nature of these distributions differs considerably from those constructed from the MIRACLE study. Nevertheless, the main features of these graphs are similar to those defined for the MIRACLE study with 2 important exceptions. First, the absolute rate of responders is slightly less than in the MIRACLE study (though we could not test whether the responder rates were statistically different between these 2 studies); and second, the difference in response rates between treatment and control groups (Fig. 4F) decreases more steeply as the value of the success criterion increases. Over the range of 1 to 2 mL $O_2 \cdot kg \cdot min$, the difference in response rate varies between $\sim$5% and 12%.

Values for 6-minute hall walk test and Minnesota living With Heart Failure Questionnaire were also provided for the MIRACLE study.[2] Again, assuming that the data are normally distributed, the estimated distributions, the rate of response as a function of the success criterion and the difference in responder rates between the 2 groups as a function of success criterion are summarized in Fig. 5. In contrast to the peak VO_2, for the 6-minute hall walk test, the difference between control and treatment groups was $\sim$18% to 19% (relatively independent of the success criterion used). This difference for the Minnesota Living With Heart Failure Questionnaire (for which a decreased value indicates a beneficial response) was $\sim$15% to 16% (also relatively independent of the selected success criterion).

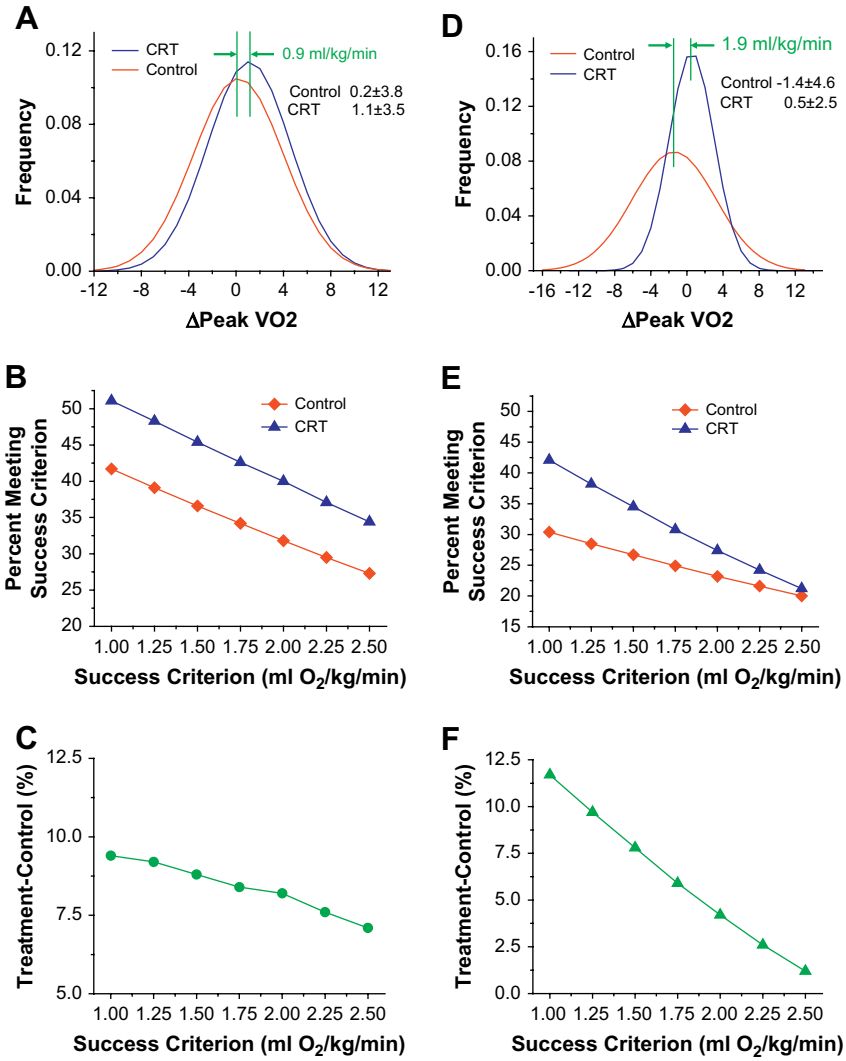

Fig. 4. Simulations of distributions of changes in peak oxygen consumption (VO_2) (A), results of a responder analysis as a function of success criterion (B) and difference in response rates between control and treatments groups as a function of success criterion (C) for the Multicenter InSync Randomized Clinical Evaluation (MIRACLE) study of cardiac resynchronization therapy (CRT)[2]. Same plots shown for RHYTHM implantable cardioverter defibrillator (ICD) study (D, E, F)[7].

Responder Analysis Applied to Original Data From a Study of CCM

The results discussed are simulations with several major assumptions. The Fix Heart Failure-4 study was multicenter, randomized, double-blind study of the impact of CCM on exercise tolerance assessed by peak VO_2 analyzed in a blinded core lab that randomized 164 patients with NYHA Class II or III symptoms who were not eligible for CRT.[8] All patients were implanted with a CCM device; half were randomized to receive CCM treatment for 3 months, whereas the device remained off in the other half. After 3 months, both groups crossed over to the opposite treatment. There was a large placebo effect during the first 3 months, so that both groups improved significantly. However, during the second 3-month period, patients switching from control to active treatment continued to show improvement and patients switching from active treatment to control returned back to baseline. The average results comparing baseline with 6 months are summarized in Fig. 6A. If we analyze these data according to the same responder analysis presented previously, we arrive at curves that are generally similar to those illustrated for the CRT studies (Fig. 6B). One significant difference, however, as shown in Fig 6C, is that as the threshold for individual patient success is increased, the difference between the control and treatment actually increases significantly to reach ~18% at a value of 2.5 mL $O_2 \cdot kg \cdot min$. Thus, in general, the qualitative aspects of the responders analysis arrived at by simulation are confirmed when applied to a real set of data.

Discussion

We have reviewed the principles of a "responder analysis" applied to simulated data from 2 studies of CRT and to

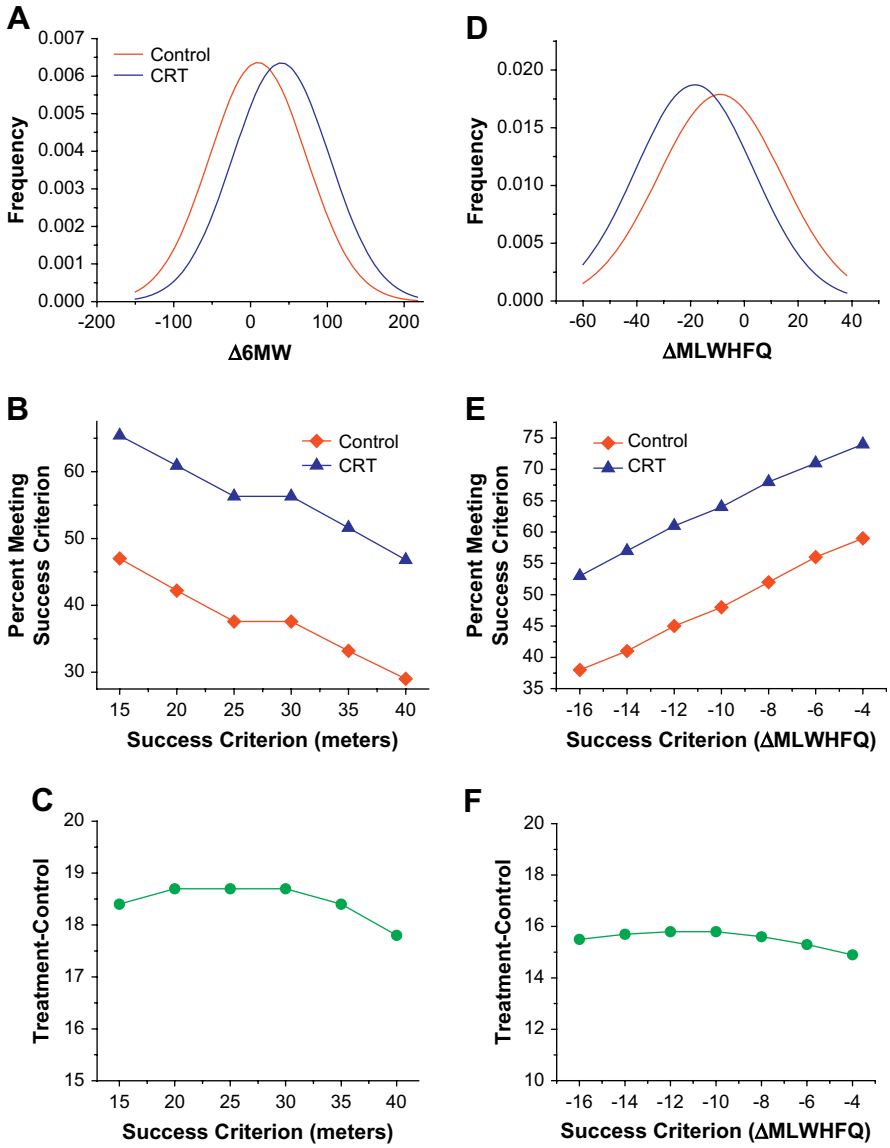

Fig. 5. Simulations of distributions of changes in 6-minute walk test (6MW, A), results of a responder analysis as a function of success criterion (B), and difference in response rates between control and treatments groups as a function of success criterion (C) for the Multicenter InSync Randomized Clinical Evaluation (MIRACLE) study of cardiac resynchronization therapy (CRT)[2]. Same plots shown for Minnesota Living with Heart Failure Questionnaire (D, E, F).

1 original set of data obtained from a recent study of CCM. The importance of this analysis stems from the fact that the cardiac device branch of the US Food and Drug Administration is encouraging sponsors to employ responder analysis as the approach for analyzing the primary and secondary efficacy end point in cases when QoL or ET measures are being used. Yet, because there is no significant direct experience with this type of analysis in a prospective study of a heart failure therapy, a review of the theory, presentation of results of simulations based on available data for a proven therapy, and presentation of results from a real set of data provides a thorough platform for open discussion of the potential pros, cons, and challenges associated with this type of analysis.

We have identified certain advantages and disadvantages to the use of a responder analysis in lieu of a difference in means analysis. A responder analysis allows one to inform a patient of his or her chance of experiencing a clinically meaningful improvement in the variable that was measured, both in absolute terms and in comparison to a control group. However, there are no objective criteria to specify the threshold value for defining individual patient success. In addition, a responder analysis and difference in means analysis may provide different results. For example, a therapy that produces a dramatic benefit in a small number of patients will look drastically different depending on what cutoff is used in a responder analysis. Other theoretical examples in which a responder analysis may lead to different

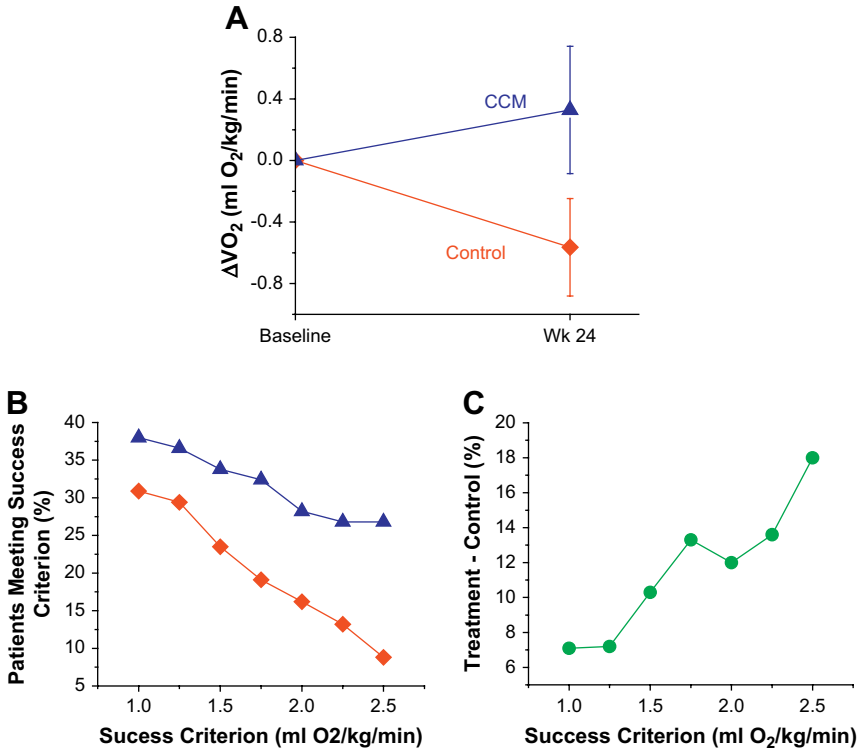

Fig. 6. (A) Changes in peak oxygen consumption (VO_2) measured in a control and treatment group from the Fix Heart Failure-4 study of cardiac contractility modulation (6). (B) Results of a responder analysis as a function of success criterion. (C) Difference in response rates between control and treatments groups as a function of success criterion.

conclusions than a comparison of means are shown in Fig. 2. Furthermore, for the examples shown previously, a responder analysis suggested CRT to be of similar or even slightly greater effectiveness in the MIRACLE study than in RHYTHM ICD, despite the fact that the mean change in peak VO_2 between control and treatment groups was more than twice as large in RHYTHM ICD. This difference was largely because of a difference in standard deviations of the changes in peak VO_2 between the control and treatment groups in RHYTHM ICD.

Another potential limitation is that there can be more uncertainty with sample size calculations and the overall design of the statistical analysis plan for a responder analysis than with a standard comparison of means analysis. Sample size estimation for a responder analysis would assume normal distributions, and deviations from normality and subtle differences between groups in standard deviations can impact significantly on the final analysis. As illustrated in some of the examples previously, there can be greater differences in response rates of control and treatment groups when a larger threshold value for success is employed (Fig. 6C), or it may turn out that this difference is largest with a lower threshold value (Fig. 4F). Thus the responder analysis can potentially be sensitive to the nature of the final distributions of results in the 2 study groups. Of course, it is emphasized that these results are *simulations*, and may only provide rough estimates of reality. Nevertheless, they point to important aspects of a responder analysis that must be better understood.

Another point for consideration is whether the threshold value for defining individual patient success should be based on an absolute or relative change from baseline of an outcome measure. The illustrations of the present report have been based on absolute changes in ET and QoL measures. However, for example, does a 1 mL $O_2 \cdot kg \cdot min$ increase in peak VO_2 have the same clinical meaning regarding efficacy in a patient with a baseline peak VO_2 value of 10 as in a patient with a baseline value of 16 mL $O_2 \cdot kg \cdot min$? Or, is it more meaningful to require a certain percent increase from baseline to define success? If the criterion for success was defined as a 10% increase from baseline, a 1 mL $O_2 \cdot kg \cdot min$ increase would be sufficient for the first patient to pass the threshold of success, but the second patient would require a 1.6 mL $O_2 \cdot kg \cdot min$ increase. If such an approach is selected, how should the percent increase be selected for any given parameter of interest? There are no objective data on how to answer such important questions.

Summary and Conclusions

The results of randomized trials form the foundation for a physician's decision to recommend use of a therapy and a patient's decision to receive treatment. In this regard, use of a responder analysis may be important to both physician and patient by specifying a patient's odds of experiencing a clinically meaningful treatment effect. However,

optimal choice of the definition for success is not obvious and could vary in different circumstances. Thus, with the introduction of a new method of analysis, new questions and uncertainties are also introduced. Of course, in the context of a clinical trial, a responder analysis and a comparison of mean changes can both be done. However, typically only 1 analysis can be declared for the primary assessment of efficacy, and it is possible that the different approaches could lead to different conclusions about the therapy (Fig. 2). One way to address many of these issues might be for investigators to routinely present graphs of the actual distributions of response in the 2 study groups (as in Figs. 4A and 4D). The differences between these distributions in the control and treatment groups, supplemented by results of appropriate statistical tests, could provide a complete characterization of both the natural history of the disease and the impact of the treatment under investigation.

As investigators, regulators, and practicing physicians seek to improve methods for testing and describing the effectiveness of therapies for chronic heart failure, the emergence of the ''responder analysis'' represents an interesting possibility. Additional prospective studies and procurement of individual patient data from prior studies of proven therapies (eg, CRT) would help advance understanding of the optimal way to use a responder analysis of QoL and ET data within the context of clinical trials of heart failure therapies.

Acknowledgments

DB is an employee of IMPULSE Dynamics, the sponsor of the FIX-HF-4 study of cardiac contractility modulation. MP, WTA and AK are consultants to Impulse Dynamics. MB participates in a speaker's bureau for IMPULSE Dynamics, Medtronics and St Jude Medical. AK is on the speaker's bureau for Medtronic, St. Jude, and receives grant support from Medtronic and St. Jude. AK is also a consultant to Lifewatch.

References

1. Hunt SA, Abraham WT, Chin MH, et al. ACC/AHA 2005 guideline update for the diagnosis and management of chronic heart failure in the adult: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Writing Committee to Update the 2001 Guidelines for the Evaluation and Management of Heart Failure): developed in collaboration with the American College of Chest Physicians and the International Society for Heart and Lung Transplantation: endorsed by the Heart Rhythm Society. Circulation 2005;112:e154–235.
2. Abraham WT, Fisher WG, Smith AL, et al. Cardiac resynchronization in chronic heart failure. N Engl J Med 2002;346:1845–53.
3. Bristow MR, Saxon LA, Boehmer J, et al. Cardiac-resynchronization therapy with or without an implantable defibrillator in advanced chronic heart failure. N Engl J Med 2004;350:2140–50.
4. Cleland JG, Daubert JC, Erdmann E, et al. Longer-term effects of cardiac resynchronization therapy on mortality in heart failure [the CArdiac REsynchronization-Heart Failure (CARE-HF) trial extension phase]. Eur Heart J 2006;27:1928–32.
5. Birnie DH, Tang AS. The problem of non-response to cardiac resynchronization therapy. Curr Opin Cardiol 2006;21:20–6.
6. Borggrefe MM, Lawo T, Butter C, et al. Randomized, double blind study of non-excitatory, cardiac contractility modulation electrical impulses for symptomatic heart failure. Eur Heart J 2008;29:1019–28.
7. Lalukota K, Cleland JG, Ingle L, Clark AL, Coletta AP. Clinical trials update from the Heart Failure Society of America: EMOTE, HERB-CHF, BEST genetic sub-study and RHYTHM-ICD. Eur J Heart Fail 2004;6:953–5.
8. Borggrefe M, Lawo T, Butter C, et al. Randomized, double blind study of non-excitatory, cardiac contractility modulation (CCM) electrical impulses for symptomatic heart failure. Eur Heart J. 2008.