Agreement Is Poor Among Current Criteria Used to Define Response to Cardiac Resynchronization Therapy

Brandon K. Fornwalt, PhD; William W. Sprague, BS; Patrick BeDell, BBA, RDCS; Jonathan D. Suever, BS; Bart Gerritse, PhD; John D. Merlino, MD; Derek A. Fyfe, MD, PhD; Angel R. León, MD; John N. Oshinski, PhD

**Background**—Numerous criteria believed to define a positive response to cardiac resynchronization therapy have been used in the literature. No study has investigated agreement among these response criteria. We hypothesized that the agreement among the various response criteria would be poor.

**Methods and Results**—A literature search was conducted with the keywords “cardiac resynchronization” and “response.” The 50 publications with the most citations were reviewed. After the exclusion of editorials and reviews, 17 different primary response criteria were identified from 26 relevant articles. The agreement among 15 of these 17 response criteria was assessed in 426 patients from the Predictors of Response to Cardiac Resynchronization Therapy (PROSPECT) study with Cohen’s \( \kappa \)-coefficient (2 response criteria were not calculable from PROSPECT data). The overall response rate ranged from 32% to 91% for the 15 response criteria. Ninety-nine percent of patients showed a positive response according to at least 1 of the 15 criteria, whereas 94% were classified as a nonresponder by at least 1 criterion. \( \kappa \)-Values were calculated for all 105 possible comparisons among the 15 response criteria and classified into standard ranges: Poor agreement (\( \kappa < 0.4 \)), moderate agreement (\( 0.4 < \kappa < 0.75 \)), and strong agreement (\( \kappa \geq 0.75 \)). Seventy-five percent of the comparisons showed poor agreement, 21% showed moderate agreement, and only 4% showed strong agreement.

**Conclusions**—The 26 most-cited publications on predicting response to cardiac resynchronization therapy define response using 17 different criteria. Agreement between different methods to define response to cardiac resynchronization therapy is poor 75% of the time and strong only 4% of the time, which severely limits the ability to generalize results over multiple studies. (Circulation. 2010;121:1985-1991.)

**Key Words:** heart failure ■ cardiac resynchronization therapy ■ pacing ■ pacemakers

Predicting whether a patient will benefit, or “respond,” to cardiac resynchronization therapy (CRT) has been the focus of more than 500 publications during the last 5 years; however, the definition of response to CRT varies widely between studies, and numerous criteria to define a positive response to CRT exist in the literature. “Echocardiographic” response is typically assessed by quantifying the change in left ventricular ejection fraction (LVEF) or left ventricular end-systolic volume (LVESV) 3 to 6 months after CRT implantation. “Clinical” response is assessed with the increase in the distance walked in 6 minutes or improvement in New York Heart Association functional class 3 to 6 months after CRT implantation. Some studies have defined response to CRT as a combination of several clinical measures or as a combination of both clinical and echocardiographic measures.
Statistics and Quantification of Agreement

The Cohen $\kappa$-coefficient was used to assess agreement between the different response criteria. The $\kappa$-coefficient is an accepted statistical coefficient that is used to assess agreement between methodologies.28,29 The $\kappa$-coefficient ranges from $-1$ (perfect disagreement) to $+1$ (perfect agreement), and a $\kappa$-coefficient of 0 indicates that the amount of agreement was exactly that expected by chance.28 A

$\kappa$-coefficient $\geq 0.75$ was defined as strong agreement, $0.4 < \kappa < 0.75$ was defined as moderate agreement, and $\kappa < 0.4$ was defined as poor agreement (Table 2).29

The $\kappa$-coefficient was calculated with the following equation:

\[
\kappa = \frac{\text{observed agreement} - \text{expected agreement}}{100 - \text{expected agreement}}
\]

Expected agreement is the percentage of cases in which 2 response criteria would agree on the basis of chance alone and can be calculated as follows:

\[
\text{expected agreement} = \frac{\text{response} \times \text{non-response}}{\text{total cases}}
\]

An example of a $\kappa$-value calculation is shown in Table 3 and presented in the Results section.

Two of the 17 response criteria24,25 could not be calculated from PROSPECT data because (1) oxygen consumption at peak exercise was not measured in PROSPECT and (2) height and weight at 6 months were not measured in PROSPECT, which precludes the

| Table 1. Seventeen Different Response Criteria Identified From the 26 Relevant Publications |
|----------------------------------|------------------|
| **Response criteria** | **Echocardiographic** |
| 1. $\uparrow$ LVEF $\geq 5\%$ (absolute) | $\uparrow$ LVEF $\geq 5\%$ |
| 2. $\uparrow$ LVEF $\geq 15\%$ | $\uparrow$ LVEF $\geq 15\%$ |
| 3. $\downarrow$ LVEF $\geq 10\%$ and did not die of progressive HF within 6 months | $\downarrow$ LVEF $\geq 15\%$ |
| 4. $\downarrow$ LVEF $> 15\%$ | $\downarrow$ LVEF $> 15\%$ |
| 5. LVEF $< 115\%$ of baseline | LVEF $< 115\%$ of baseline |
| 6. $\downarrow$ LVESVI $> 15\%$ | $\downarrow$ LVESVI $> 15\%$ |
| 7. $\downarrow$ LVEDV $> 15\%$ | $\downarrow$ LVEDV $> 15\%$ |
| 8. $\uparrow$ Stroke volume $\geq 15\%$ | $\uparrow$ Stroke volume $\geq 15\%$ |

**Clinical**

9. $\downarrow$ NYHA $\geq 12,13,14$
10. $\downarrow$ NYHA $\geq 1$ and did not die of progressive HF within 6 months23
11. $\downarrow$ NYHA $\geq 1$ and $\uparrow$ 6MWD $\geq 25\%$15
12. $\downarrow$ NYHA $\geq 1$ and $\uparrow$ 6MWD $\geq 25\%$ and did not die of progressive HF within 6 months16,17
13. $\uparrow$ 6MWD $> 10\%$, no heart transplant, did not die of progressive HF within 6 months11
14. ($\downarrow$ NYHA $\geq 1$ or $\uparrow$ VO$_2$max $> 10\%$ or $\uparrow$ 6MWD $> 10\%$) and alive, no hospitalization for decompensated HF24
15. Two of 3:5
   - $\downarrow$ NYHA $\geq 1$
   - $\uparrow$ 6MWD $\geq 50$ m
   - $\downarrow$ QOL $\geq 15$
16. Clinical composite score improved$^10$

**Combined**

17. ($\uparrow$ LVEF $\geq 5\%$ [absolute] or $\uparrow$ 6MWD $\geq 30$ m) and ($\downarrow$ NYHA $\geq 1$ or $\downarrow$ QOL $\geq 10$)$^{18}$

$\uparrow$ Indicates increase; LVEF, left ventricular ejection fraction; $\downarrow$, decrease; HF, heart failure; LVEF, left ventricular end-systolic volume; LVESVI, LVESV indexed by body surface area; LVEDV, left ventricular end-diastolic volume; NYHA, New York Heart Association functional class; 6MWD, 6-minute walk distance; VO$_2$max, oxygen consumption at peak exercise; and QOL, quality-of-life score.

If the authors did not specify whether death was considered a nonresponse, then it was assumed that deaths were excluded.
ability to calculate the LVESV index. Agreement among the 15 remaining response criteria was assessed by calculating a κ-value for all possible pairs of criteria. Thus, 105 κ-values were calculated. Group κ-values were then quantified (mean, median, and range) to summarize the agreement among criteria from 4 different groups: (1) All 105 comparisons, (2) comparisons between 2 echocardiographic criteria, (3) comparisons between 2 clinical criteria, and (4) comparisons between an echocardiographic criterion and a clinical criterion. Mean group κ-values were compared with a permutation test, and a Bonferroni correction was applied to account for multiple comparisons. To justify that the sample size to estimate κ-values was large enough, a bootstrap resampling procedure was used with a Kolmogorov-Smirnov test to assess the normality of the resulting distribution. P<0.05 was defined as statistically significant.

Subgroup Analysis Excluding Response Criteria Quantified Short Term and at 3 Months

One response criterion (increase in stroke volume ≥15%) was used in the literature as a short-term response measure that was quantified within 2 days of CRT implantation. In addition, 2 response criteria from the literature were only used in studies that assessed response at 3 months (criterion 15 and 17 in Table 1). Short-term (<2 days) and 3-month data for calculating all 15 response measures were not collected in PROSPECT, so all criteria were assessed with data from the 6-month visit. To ensure that the present results were not confounded by calculating these short-term and 3-month measures from 6-month follow-up data, we performed a subgroup analysis in which we excluded them and recalculated the group κ-values.

Results

Response Rates

The percentage of patients defined as having a positive “response” to CRT ranged from 32% to 91% for the 15 response criteria (Table 4). All 15 criteria could be calculated in 250 of the 426 patients in PROSPECT. Of these 250 patients, 99% showed a positive response according to at least 1 of the 15 criteria, whereas 94% were classified as nonresponders by at least 1 criterion. Similarly, 95% of patients showed a positive response by at least 2 of the 15 criteria, whereas 87% were classified as nonresponders by at least 2 criteria.

Example Calculation of κ-Value

An example of the κ-value calculation between response criterion 3 (decrease in LVESV ≥10%, no death due to heart failure) and response criterion 13 (increase in 6-minute walk distance ≥10%, no heart failure death, no transplant) is given in Table 3. Equation 1 then shows that

\[ \kappa = \frac{25\% - 0.38}{1 - 0.38} \]

The observed agreement was 0.39 + 0.15 = 54%. Equation 1 then shows that \( \kappa = 0.02 \) for Table 3, which suggests poor agreement after accounting for the level of agreement expected due to chance.

Agreement Among the 15 Response Criteria

The 15 response criteria showed poor agreement as a group (Figures 2 and 3; mean \( \kappa = 0.22 \pm 0.24 \), median = 0.14, range = −0.2 to 0.97). Seventy-nine (75%) of the 105 κ-values were classified as having poor agreement, whereas 22 κ-values (21%) were classified as having moderate agreement (Figure 3). Only 4 pairs of response criteria of the 105 total pairs (4%) showed strong agreement, and 2 of these 4 pairs were comparisons between a response criterion that excluded mortality and the same exact criterion that defined death as a nonresponse. The 7 echocardiographic response criteria also showed poor agreement among each other (Figure 4; mean \( \kappa = 0.35 \pm 0.28 \), median = 0.29, range = −0.2 to 0.88). The 7 clinical response criteria showed moderate agreement (Figure 4; mean \( \kappa = 0.44 \pm 0.23 \), median = 0.43,
range = 0.14 to 0.97). Agreement between echocardiographic and clinical criteria was poor (Figure 4; mean $\kappa = 0.05 \pm 0.05$, median = 0.04, range = -0.03 to 0.17), with all 49 $\kappa$-values showing poor agreement. The agreement among the echocardiographic parameters was not significantly different from the agreement among the clinical parameters (uncorrected $P = 0.35$). The response criterion based on a combination of both echocardiographic and clinical measures showed significantly better agreement ($P = 0.003$) with clinical response criteria (0.44 $\pm$ 0.10) than with echocardiographic response criteria (0.21 $\pm$ 0.12). Bootstrap resampling of the $\kappa$-statistic comparing clinical composite response and a 15% reduction in LVEF justified that the sample size was large enough to estimate the $\kappa$-value (Kolmogorov-Smirnov test for normality $P > 0.15$).

### Subgroup Analysis Excluding Response Criteria Quantified Short Term and at 3 Months

Exclusion of the short-term and 3-month response criteria did not significantly affect the results. After exclusion of 1 short-term (criterion 8 in Table 1) and two 3-month (criteria 15 and 17 in Table 1) response measures, the agreement among the 12 remaining response criteria was poor as a group (mean $\kappa = 0.22 \pm 0.26$, median = 0.14, range = -0.03 to 0.97). Agreement among the 6 remaining echocardiographic response criteria was moderate (mean $\kappa = 0.44 \pm 0.24$, median = 0.47, range = 0.16 to 0.88). Agreement among the 6 remaining clinical response criteria was also moderate (mean $\kappa = 0.42 \pm 0.27$, median = 0.32, range = 0.14 to 0.97). Finally, agreement between echocardiographic and clinical criteria remained poor (mean $\kappa = 0.05 \pm 0.05$, median = 0.04, range = -0.03 to 0.17).

### Discussion

The major findings of this study are as follows: (1) The 26 most-cited publications on predicting response to CRT used 17 different primary response criteria, and the level of agreement not due to chance among 15 of these response criteria was poor 75% of the time and strong only 4% of the time.

Figure 2. Agreement among the 15 response criteria was poor. The $\kappa$-axis shows the following ranges delineated by dotted lines: strong agreement ($\kappa \geq 0.75$), moderate agreement ($0.4 < \kappa < 0.75$), and poor agreement ($\kappa < 0.4$). The worst agreement was between echocardiographic (Echo) and clinical (Clin) parameters. *$P < 0.001$ vs “Echo vs Echo” and “Clin vs Clin.”

Figure 3. Agreement among the 15 response criteria was classified as poor for 75% of the 105 possible comparisons. $\kappa$-Values are color-coded according to the following ranges: green = strong agreement ($\kappa \geq 0.75$), yellow = moderate agreement ($0.4 < \kappa < 0.75$), and red = poor agreement ($\kappa < 0.4$). Echo indicates echocardiographic response criteria; Clin, clinical response criteria.
in the 426 patients enrolled in the PROSPECT study; (2) agreement between echocardiographic and clinical response criteria was poor and nearly equal to the level of agreement expected by chance; (3) the percentage of patients defined as having a positive response to CRT ranged from 32% to 91% for the 15 response criteria; and (4) 99% of patients were classified as a responder by at least 1 of the 15 criteria, whereas 94% were classified as a nonresponder by at least 1 criterion.

Comparison to the Literature

To the best of our knowledge, no study has quantified agreement among response criteria with a $\kappa$-coefficient; however, a recent study by Bleeker et al.2 aimed to quantify the agreement between echocardiographic and clinical measures of response to CRT. The authors compared a decline in New York Heart Association class (clinical response) with a 15% decrease in LVESV (echocardiographic response) in 144 consecutive patients undergoing CRT. The authors concluded that “the agreement between [clinical response and echocardiographic response] was good” based on the observed agreement of 76%. However, their data show that clinical and echocardiographic responses would be expected to agree 52% of the time on the basis of chance alone. The study did not calculate $\kappa$-values to account for this expected level of agreement due to chance. We estimated the $\kappa$-value to be 0.50 from their data, which is higher than the value of 0.17 observed in the present study. However, the main conclusion that should be drawn from both studies is similar: The agreement between echocardiographic and clinical criteria for defining a positive response to CRT is only slightly better than that expected by chance alone.

In the MIRACLE trial (Multicenter InSync Randomized Clinical Evaluation), correlation between the change in left ventricular end-diastolic volume and change in New York Heart Association class after 6 months of CRT was weak ($r=0.13$).30 In addition, the correlation between the change in distance walked in 6 minutes and change in left ventricular ejection fraction was weak ($r=0.15$).30 These data are consistent with the present results, which show poor agreement between clinical and echocardiographic response criteria.

Previous studies have reported different rates of response to CRT when different definitions of response are used within the same population. For example, the PROSPECT study reported that 56% of patients were echocardiographic responders (defined by a reduction in LVESV of at least 15%), whereas 69% of patients were clinical responders (defined by an improvement in the clinical composite score).10 However, the actual response rate does not tell the entire story: Table 3 shows 2 different response criteria with identical response rates of 62%, and despite the identical response rates, the criteria show very poor agreement ($\kappa=0.02$). Thus, assessment of agreement with the $\kappa$-statistic provides valuable information in addition to the overall response rate of the population.

---

Figure 4. Agreement among the response criteria was poor 75% of the time and strong only 4% of the time. $\kappa$-Values are color-coded according to the following ranges: Green = strong agreement ($\kappa>0.75$), yellow = moderate agreement (0.4 < $\kappa<0.75$), and red = poor agreement ($\kappa<0.4$). LVEF indicates left ventricular ejection fraction; HF, heart failure; NYHA, New York Heart Association; LVESV, left ventricular end-systolic volume; LVEDV, left ventricular end-diastolic volume; 6MWD, distance walked in 6 minutes; and QOL, quality-of-life score.
Other Inconsistencies in Defining Response to CRT

Length of Follow-Up

Another area of inconsistency in defining response to CRT is the length of the follow-up period after which a patient is deemed either a responder or a nonresponder. Some studies focused on short-term (1 to 2 days) response, whereas most focused on 3-month or 6-month response. CRT has been shown to have persistent, increasing benefits with a longer mean follow-up period of 29.5 months. We defined response at 6 months because this was the prespecified follow-up period for the PROSPECT study. We also performed a subgroup analysis after excluding criteria that were assessed in the literature at short-term and 3-month follow-up only, and this did not change the present results. Future studies will be needed to address agreement among the different lengths of follow-up.

Mortality

Whether death should be considered a nonresponse to CRT is another area in which there is inconsistency. There are at least 3 different methods that authors have used to incorporate death into their response criteria: (1) Death due to worsening heart failure is included in the nonresponder group, and (2) death due to any cause is included in the nonresponder group, and (3) deaths are excluded from analysis. Numerous publications fail to specify how death was incorporated into response criteria despite enrolling consecutive patients and following them for a 3- to 6-month period. Although inclusion of all-cause mortality as a criterion for nonresponse may not be appropriate, a patient who dies of progressive heart failure is included in the nonresponder group, (2) death due to any cause is included in the nonresponder group, and (3) deaths are excluded from analysis. Moreover, numerous publications fail to specify how death was incorporated into response criteria despite enrolling consecutive patients and following them for a 3- to 6-month period. Although inclusion of all-cause mortality as a criterion for nonresponse may not be appropriate, a patient who dies of progressive heart failure should, objectively, be classified as a nonresponder. Regardless, there is no consistent method for incorporating mortality into the definition of response to CRT, and this needs to be standardized.

A Consensus Definition of “Response to CRT”

Because heart failure is a debilitating life-threatening disease, an effective heart failure therapy should treat both symptoms and quality and duration of life. Thus, measures of “response” to CRT should either directly measure outcomes or have a surrogate relationship with benefits in heart failure symptoms, quality of life, and duration of life. The clinical composite score is a measure of response that accounts for all of these factors and may be the best overall choice for defining response in future CRT trials.

Study Limitations

The results of the present study are limited by the fact that we used data from a single study (PROSPECT). However, PROSPECT was a multicenter study that enrolled 457 wide-QRS patients from 53 different centers across Europe, Hong Kong, and the United States. We would expect similar results from other large, multicenter databases.

The present results show that many different methods to define a positive response to CRT are being used in the literature and show poor agreement among each other. This begs the question, which method should we use in the future to determine whether a patient benefited from CRT? The present study did not attempt to address this question, and future studies will need to explore this important issue.

The definition of clinically acceptable agreement based on the k-coefficient is not standardized. Fleiss proposed a threshold of ±0.75 to define strong evidence for agreement that is not due to chance, which is what we used; however, this threshold is somewhat arbitrary. Landis and Koch proposed that any value of k above 0.60 suggests substantial agreement, and k>0.80 implies “almost perfect” agreement. However, the use of a different threshold, such as 0.6, to define strong agreement would not significantly alter the present results; 4 of the 105 κ-statistics that we calculated were greater than 0.8, and only 8 were ≥0.6. Thus, regardless of the threshold used, we observed mostly poor agreement among the 15 different response criteria.

Conclusions

The 26 most-cited publications on predicting response to CRT define response using 17 different criteria. Agreement between these different published methods to define response to CRT is poor 75% of the time and strong only 4% of the time. This inconsistency in the definition of response to CRT severely limits the ability to generalize results over multiple studies and hinders progress in the field.

Sources of Funding

This work was supported by grants from the American Heart Association (Grant-in-Aid No. 0855386E) and the National Institutes of Health (HL089160) to Dr Oshinski. Dr Fornwalt was supported in part by National Institutes of Health training grant No. 5 T32 GM008169.

Disclosures

Dr Gerritse is an employee of Medtronic Inc and owns company stock. The remaining authors report no conflicts.

References

8. Yu CM, Zhang Q, Chan YS, Chan CK, Yip GKW, Kum LCC, Wu EB, Lee FW, Lam YY, Chan S, Fung JWH. Tissue Doppler velocity is superior to displacement and strain mapping in predicting left ventricular
Poor Agreement Among CRT Response Criteria


Go to http://cme.ahajournals.org to take the CME quiz for this article.
Agreement Is Poor Among Current Criteria Used to Define Response to Cardiac Resynchronization Therapy

Brandon K. Fornwalt, William W. Sprague, Patrick BeDell, Jonathan D. Suever, Bart Gerritse, John D. Merlino, Derek A. Fyfe, Angel R. León and John N. Oshinski

_Circulation_. 2010;121:1985-1991; originally published online April 26, 2010; doi: 10.1161/CIRCULATIONAHA.109.910778

_Circulation_ is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2010 American Heart Association, Inc. All rights reserved.
Print ISSN: 0009-7322. Online ISSN: 1524-4539

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://circ.ahajournals.org/content/121/18/1985

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in _Circulation_ can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to _Circulation_ is online at:
http://circ.ahajournals.org//subscriptions/