Validation of Exercise Capacity as a Surrogate Endpoint in Exercise-Based Rehabilitation for Heart Failure
A Meta-Analysis of Randomized Controlled Trials

Oriana Ciani, PhD,a,b Massimo Piepoli, MD,c Neil Smart, PhD,d Jamal Uddin, MSc,e,f Sarah Walker, PhD,g Fiona C. Warren, PhD,a Ann D. Zwisler, MD,f,g Constantinos H. Davos, PhD,h Rod S. Taylor, PhD,a,f,g

ABSTRACT

OBJECTIVES This study sought to validate exercise capacity (EC) as a surrogate for mortality, hospitalization, and health-related quality of life (HRQOL).

BACKGROUND EC is often used as a primary outcome in exercise-based cardiac rehabilitation (CR) trials of heart failure (HF) via direct cardiorespiratory assessment of maximum oxygen uptake (VO2peak) or through submaximal tests, such as the 6-min walk test (6MWT).

METHODS After a systematic review, 31 randomized trials of exercise-based CR compared with no exercise control (4,784 HF patients) were included. Outcomes were pooled using random effects meta-analyses, and inverse variance weighted linear regression equations were fitted to estimate the relationship between the CR on EC and all-cause mortality, hospitalization, and HRQOL. Spearman correlation coefficient (\(r\)), \(R^2\) at trial level, and surrogate threshold effect (STE) were calculated. STE represents the intercept of the prediction band of the regression line with null effect on the final outcome.

RESULTS Exercise-based CR is associated with positive effects on EC measured through VO2peak (\(\pm 3.10\) ml/kg/min; 95% confidence interval [CI]: 2.01 to 4.20) or 6MWT (\(\pm 41.15\) m; 95% CI: 16.68 to 65.63) compared to control. The analyses showed a low level of association between improvements in EC (VO2peak or 6MWT) and mortality and hospitalization. Moderate levels of correlation between EC with HRQOL were seen (e.g., \(R^2 < 52\%\); \(|r| < 0.72\)). Estimated STE was an increase of 5 ml/kg/min for VO2peak and 80 m for 6MWT to predict a significant improvement in HRQOL.

CONCLUSIONS The study results indicate that EC is a poor surrogate endpoint for mortality and hospitalization but has moderate validity as a surrogate for HRQOL. Further research is needed to confirm these findings across other HF interventions. (J Am Coll Cardiol HF 2018;6:596–604) © 2018 The Authors. Published by Elsevier on behalf of the American College of Cardiology Foundation. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Enhancement of exercise capacity (EC) is a key aspect of the lifestyle and management of patients with heart failure (HF) (1,2). The gold standard approach for measuring EC is a maximal (or symptom-limited) exercise test with direct cardiorespiratory assessment of peak oxygen uptake (VO2peak) or via indirect submaximal tests, including the 6-min walk test (6MWT) (3).

EC is often used as a primary outcome in HF trials and is accepted by the United States Food and Drug

From the "Institute of Health Research, University of Exeter Medical School, Exeter, United Kingdom;" Centre for Research on Health and Social Care Management, Bocconi University, Milan, Italy; "Heart Failure Unit, Guglielmo da Saliceto Hospital, Piacenza, Italy;" School of Science and Technology, University of New England, Armidale, Australia; "Department of Cardiac Surgery, Ibrahim Cardiac Hospital & Research, Institute, Dhaka, Bangladesh;" National Institute of Public Health, University of Southern Denmark, Copenhagen, Denmark; "Danish Knowledge Centre for Rehabilitation and Palliative Care, University Hospital Odense and University of Southern Denmark, Odense, Denmark;" and the "Cardiovascular Research Laboratory, Biomedical Research Foundation Academy of Athens, Athens, Greece. Dr. Ciani is funded by a post-doctoral scholarship from the University of Exeter Medical School (Exeter, United Kingdom). All authors have reported that they have no relationships relevant to the contents of this paper to disclose.

Manuscript received February 22, 2018; accepted March 27, 2018.

ISSN 2213-1779
Administration (4). Although it is a measure of function and clinical benefit, EC is a surrogate endpoint rather than a final patient-relevant outcome, such as mortality, hospital admission, or health-related quality of life (HRQOL).

For a surrogate endpoint to be considered a valid surrogate endpoint (i.e., an adequate substitute for the final outcome), several levels of evidence must be provided (5). First, there needs to be biological plausibility of the relationship between the surrogate and the final outcome. Second, observational or epidemiological studies are required to show a consistent association between the surrogate and the final outcome. Third, the treatment effect on the surrogate must correspond with the treatment effect on the final outcome. Are observational epidemiological studies are required to show a consistent association between the intervention effect of CR in HF on mortality, hospitalization, or quality of life (HRQOL).

Screening of full study reports was undertaken by 1 of the authors (O.C., J.U.) and checked by a second author (R.S.T.).

**DATA EXTRACTION.** For each study, we extracted the following information: first author, publication year, geographical location, sample size and ratio of intervention to control, study follow-up duration, setting of exercise training intervention (center- or home-based exercise program), age (mean), sex (percentage male), left ventricular ejection fraction (mean), and New York Heart Association functional class of the patient population. The continuous outcomes of EC and HRQOL were extracted at baseline and at the latest reported follow-up as mean ± SD for both exercise and control groups. The binary outcomes of mortality and hospitalization were extracted as the number of patient events at the latest follow-up relative to the number of patients randomized to each group. Whenever necessary, an online digitizer (WebPlotDigitizer) was used. Missing SDs at baseline or follow-up were imputed from confidence intervals, interquartile ranges, or SEs, and missing SDs for follow-up minus baseline change were estimated using Cochrane Handbook recommended methods (9). For trials with more than 1 exercise intervention arm, we followed the Cochrane Handbook’s approach for combining groups (9). All data was first extracted by 1 of the authors (O.C.) and then checked by another (R.S.T.). Trial quality was assessed using the Cochrane risk of bias tool (10).

**STATISTICAL ANALYSIS.** Binary outcomes at the latest follow-up were expressed as an odds ratio (OR), where OR <1.0 indicated a beneficial effect of exercise-based CR compared with control. Given the variation in reporting of outcomes across studies, we first expressed the between-group difference in EC and HRQOL for each study as a standardized mean difference and 95% confidence interval (CI). For the subset of studies that reported HRQOL using the Minnesota Living With Heart Failure (MLwHF)
questionnaire, weighted mean differences were calculated, where a negative between group mean difference (exercise minus control) in MLwHF indicates a greater HRQOL score in the training group compared to the control group. In the subset of studies that reported VO2peak (in ml/kg/min) or 6MWT (in meters), we calculated the mean difference between groups and 95% CI in original units. A positive mean difference (exercise minus control) indicated greater EC in the training group compared to the control group.

Statistical heterogeneity was assessed using the I² statistic (9), and small-study effects and publication bias were assessed using the Egger test or the Peters test. All outcomes were pooled using a DerSimonian and Laird random effects model with continuity correction when needed.

We calculated commonly reported indicators of surrogate validation (11). The correlation coefficient (r) and the R² for the relationship between treatment effect differences in EC and each of the final outcomes individually were estimated using weighting by the inverse of the variance (for the treatment effect on final outcomes). Where possible, the surrogate threshold effect (STE) was calculated. The STE represents the intercept of the prediction band of the regression line with zero effect on the final outcome (12).

We performed sensitivity analyses to assess whether our findings changed when: 1) excluding the largest included trial (HF-ACTION [Heart Failure: A Controlled Trial Investigating Outcomes of Exercise Training]) (13); 2) limiting our analysis to studies of HF with reduced ejection fraction only; or 3) limiting our analysis to studies at low risk of bias as assessed by random sequence generation (14) and allocation concealment (15).

All data analyses were conducted using Stata version 14.2 (Stata Corp., College Station, Texas) software.

RESULTS

STUDY SELECTION AND CHARACTERISTICS. A total of 31 studies and 32 comparison groups (1 study had 2 exercise intervention arms [16]) were included for analysis (Figure 1).

A summary of included studies is given in Table 1 (details are listed in Online Table 1 with supplemental references listed). The nature of exercise
training varied across studies with regard to the frequency, duration, and intensity of exercise (Online Table 1).

**RISK OF BIAS.** The overall risk of bias assessment was moderate, although several studies failed to give sufficient details to assess risk of bias criteria (Online Table 2). Reporting was found to be considerably better in more recently published studies.

**IMPACT OF EXERCISE TRAINING.** Exercise capacity. EC was reported at follow-up times ranging from 1 to 14 months. In the subset of studies reporting $V_{O2,peak}$, there was a larger increase in mean pooled EC with exercise-based CR compared with control (3.10 ml/kg/min; 95% CI: 2.01 to 4.20; 22 studies; $p < 0.001$; $I^2 = 96.6\%$) (Table 2, Online Figure 1). A similar positive finding for the exercise-based CR arm was seen in the subset of studies reporting 6MWT (41.15 m; 95% CI: 16.68 to 65.63; 10 studies; $p = 0.001$; $I^2 = 85.1\%$) (Online Figure 2).

**Final patient-relevant outcomes.** There was no difference in pooled mortality between exercise-based CR and control (OR: 0.85; 95% CI: 0.71 to 1.01; 26 trials; $p = 0.066$; $I^2 = 0\%$) (Online Figure 3). The risk of all-cause hospitalization was reduced in the exercise group compared with control (OR: 0.64; 95% CI: 0.44 to 0.93; 20 studies; $p = 0.02$; $I^2 = 60.3\%$) (Table 3, Online Figure 4).

**HRQOL at baseline and follow-up.** was reported in 21 comparisons, of which 14 used the disease-specific measure, the MLWhF questionnaire (17). Other HRQOL questionnaires reported were the Kansas City Cardiomyopathy Questionnaire (13,18), Icelandic Quality of Life (19), the Chronic Heart Failure Questionnaire (20,21), the Likert scale for symptoms (22), and the patients’ global assessment of change in quality of life (23). Across all HRQOL outcome measures, the level of HRQOL at follow-up was higher with exercise-based CR compared with control (standardized mean difference: $-0.48$ SDU; 95% CI: $-0.73$ to $-0.24$; 21 comparisons; $p < 0.0001$; $I^2 = 89.5\%$) (Online Figure 5). When pooling the subgroup of trials that reported MLWhF follow-up scores, the level of HRQOL was higher in the exercise-based CR group ($-7.24$; 95% CI: $-11.84$ to $-2.63$; 14 comparisons; $p = 0.002$; $I^2 = 67.9\%$) (Online Figure 6). All patient-relevant outcomes were measured over follow-up times from a minimum of 3 months up to 120 months.

**Evaluation of EC as surrogate endpoint.** Regression coefficients of determination ($R^2$) and correlation coefficients ($\rho$) between the change in EC and mortality or hospitalization were relatively low ($R^2 = 28\%$; $|\rho| < 0.53$) (Table 4). The coefficients for the slope of the regression line were not significantly different from zero ($p > 0.05$), confirming no clear associations between CR intervention effect on EC and clinical outcomes.

Higher correlations were seen between change in EC and HRQOL, for example, $R^2 = 32\%$ and $\rho = -0.57$ for the change in 6MWT and MLWhF; $R^2 = 52\%$; $\rho = -0.72$ for the change in $V_{O2,peak}$ and all HRQOL measures. Negative correlation coefficients indicate that larger CR effects on EC are associated with larger CR effects on HRQOL (Table 4).
**Surrogate threshold effect.** Based on the correlation analyses, we estimated STEs for EC and HRQOL. For the subset of studies reporting VO2peak (Figure 2), we estimated that an average improvement of 5 ml/kg/min in VO2peak exercise-based CR versus control is needed to predict a favorable improvement in HRQOL with exercise-based CR compared to control. For the subset of studies reporting 6MWT (Figure 3), we estimated an STE of 80 m for 6MWT to predict a significant improvement in MLwHF with exercise-based CR versus control. This pattern of link between EC and final outcomes was consistent across all sensitivity analyses (Online Tables 3 to 6).

**DISCUSSION**

Using trial-level data from RCTs of exercise-based CR for HF, we formally evaluated the evidence for EC as a surrogate endpoint for the final outcomes of mortality, hospitalization, and HRQOL. Our results show an increase in VO2peak or 6MWT with exercise-based CR to be associated with improvements in clinical outcomes (24). However, the observed levels of association indicate EC is a poor surrogate endpoint for mortality and hospitalization and has moderate validity for HRQOL. We found an STE of 5 ml/kg/min and 6MWT of 80 m. Thus, exercise-based CR would need to increase VO2peak and 6MWT by this level (or more) to have 95% confidence interval to be able to demonstrate significant improvement in HRQOL in a future trial. Sensitivity analyses confirmed VO2peak to have moderate surrogate validity for HRQOL.

**IMPLICATIONS FOR PRACTICE.** Our study findings have significant implications for future HF trials. Contrary to epidemiological (observational) evidence, our results show that intervention effects on EC are not predictive of treatment effects on the clinical events of mortality or hospitalization. However, we also showed that improvements in EC, if large enough, can be predictive of important gains in patient HRQOL.

Surrogate endpoints generally accrue more quickly than final endpoints, thus allowing for RCTs with

---

**TABLE 3** Pooled All-Cause Mortality, Hospitalization, and HRQOL (Either MLwHF or All Scales)*

<table>
<thead>
<tr>
<th>Mortality (All-Cause)</th>
<th>Hospitalization (All-Cause)</th>
<th>MLwHF (Follow-Up)</th>
<th>HRQOL (All Scales)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of studies</td>
<td>26</td>
<td>14</td>
<td>21</td>
</tr>
<tr>
<td>Effect estimate (95% CI)</td>
<td>OR: 0.85 (0.71 to 1.01)</td>
<td>OR: 0.64 (0.44 to 0.93)</td>
<td>WMD: −7.24 (−11.84 to −2.63)</td>
</tr>
<tr>
<td>p value</td>
<td>0.066</td>
<td>0.020</td>
<td>0.002</td>
</tr>
<tr>
<td>I²</td>
<td>0%</td>
<td>60.3%</td>
<td>67.9%</td>
</tr>
</tbody>
</table>

*DerSimonian and Laird random effects method. Treatment or lower than 1.0 favors exercise intervention; negative SMD/WMD favors exercise intervention. Positive weighted (by inverse variance) changes in exercise capacity values favor the exercise rehabilitation group. Number of studies, effect estimates (95% CI), p value for significance of effect size, I² statistic, and Peters or Egger test p value for small-study effect reported.

**TABLE 4** Surrogacy Validity of VO2peak and 6MWT Vs. Mortality, Hospitalization, and HRQOL Measures*

<table>
<thead>
<tr>
<th>Exercise Capacity (Follow-Up Minus Baseline Change)</th>
<th>VO2peak</th>
<th>6MWT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mortality (all-cause, log OR)</td>
<td>Coeff (SE)</td>
<td>−0.19 (0.08)</td>
</tr>
<tr>
<td></td>
<td>p value</td>
<td>0.035</td>
</tr>
<tr>
<td></td>
<td>R²</td>
<td>28%</td>
</tr>
<tr>
<td></td>
<td>p</td>
<td>−0.53</td>
</tr>
<tr>
<td>Hospitalization (all-cause, log OR)</td>
<td>Coeff (SE)</td>
<td>−0.17 (0.12)</td>
</tr>
<tr>
<td></td>
<td>p value</td>
<td>0.181</td>
</tr>
<tr>
<td></td>
<td>R²</td>
<td>12%</td>
</tr>
<tr>
<td></td>
<td>p</td>
<td>−0.35</td>
</tr>
<tr>
<td>MLwHF (follow-up value)</td>
<td>Coeff (SE)</td>
<td>−1.59 (1.64)</td>
</tr>
<tr>
<td></td>
<td>p value</td>
<td>0.388</td>
</tr>
<tr>
<td></td>
<td>R²</td>
<td>19%</td>
</tr>
<tr>
<td></td>
<td>p</td>
<td>−0.43</td>
</tr>
<tr>
<td></td>
<td>STE</td>
<td>80 m</td>
</tr>
<tr>
<td>HRQOL (all scales, follow-up value)</td>
<td>Coeff (SE)</td>
<td>−0.22 (0.07)</td>
</tr>
<tr>
<td></td>
<td>p value</td>
<td>0.013</td>
</tr>
<tr>
<td></td>
<td>R²</td>
<td>52%</td>
</tr>
<tr>
<td></td>
<td>p</td>
<td>−0.72</td>
</tr>
<tr>
<td></td>
<td>STE (ml/kg/min)</td>
<td>5</td>
</tr>
</tbody>
</table>

*Coeff (SE) and related p value for the slope of the regression line and R² derived from least-squares linear regressions weighted by the inverse of variance of the treatment effect on final outcomes. Weighted correlation coefficient (p) is shown. STE calculated as the value of the treatment effect on the surrogate outcome at which the linear regression prediction line crosses the null effect line for the treatment effect on patient-relevant outcomes. Only trials contributing to the within-study paired analysis were considered.

Coeff = coefficient; STE = surrogate threshold effect; other abbreviations as in Tables 2 and 3.
shorter follow-up periods and smaller sample sizes. Reducing trial sample size and duration ensures faster patient access to new therapies and means that trials are less expensive, which makes surrogate endpoints attractive to manufacturers and research sponsors alike. However, it is important that surrogate endpoints be carefully validated, as evidenced by the dramatic failures of surrogates used for regulatory purposes.

Our results are consistent with the review by Ferreira et al. (4), which recommended the use of EC as outcome in HF trials based on qualitative analysis of results from HF trials showing an improvement in 6MWT was associated with a favorable treatment effect of morbidity and mortality. This analysis noted that an increase of 30 to 50 m in 6MWT has been used in cardiac resynchronization therapy trials in order to gain pre-market approval. This magnitude of improvement in EC is somewhat smaller than our estimated STE of 80 m in 6MWT. Similarly, our finding of an STE of 5 ml/kg/min in VO_2peak is considerably larger than the increase of 6% in VO_2peak (i.e., ~1.0 ml/kg/min) needed to predict an improvement in the primary outcome (time to all-cause mortality or all-cause hospitalization) reported by Swank et al. (25) based on HF-ACTION. Although based on patient-level data analysis from a large RCT of exercise-based CR (HF-ACTION), this later analysis does not take into account the role of treatment; therefore, it can be used to establish the prognostic validity of VO_2peak but not the association between treatment effects on VO_2peak and treatment effects on patient-relevant outcomes as measured across a number of RCTs. In contrast, and in accordance with contemporary recommendations for surrogate endpoint validation (26), the present study derived STE values from the prediction interval around the regression line rather than from the 95% CI based on a meta-analysis of RCTs of exercise-based CR in HF.

**STUDY LIMITATIONS.** First, the methods of EC assessment varied considerably across studies. Whereas a number of studies reported VO_2peak, a small proportion directly measured VO_2peak using cardiorespiratory testing, and others predicted VO_2peak using a submaximal exercise. However, this limitation does not apply to the subgroup of studies that assessed EC using 6MWT. In addition, although
most studies reported EC at baseline and follow-up, they did not report the SD of the baseline-follow up change in EC. For these studies, we imputed the SD of change (using a correlation coefficient 0.74 from included trials) (9). Both of these factors might have introduced measurement error and masked an underlying association between EC with final outcomes. Second, there was considerable heterogeneity in exercise-based intervention applied in the included RCTs: a wide range in exercise training dose and variation in whether trials included educational and psychological co-interventions. Here, we would argue that such heterogeneity is implicit in a systematic review and meta-analysis of a complex intervention such as exercise-based CR (27). However, it could be argued that restricting our analyses to exercise-based CR trials limits the generalizability of our results. However, it is recommended that the surrogate validation be undertaken in trials across the same intervention (5). Third, a number of included RCTs studies had methodological issues or poor reporting that may have resulted in their high risk of bias. However, reassuringly, our findings were consistent when limited to the subgroup of trials of low risk of bias. Fourth, despite excluding trials with follow-up <6 months, the timing of longest assessment across studies varied from 1 to 14 months for EC and from 3 months to 10 years for final outcomes. Unfortunately, individual studies did not consistently report EC and final outcomes at repeated time points, so we could not explore this issue.

CONCLUSIONS

This systematic review and meta-analysis of RCTs of exercise-based CR shows that treatment effect on EC is a poor surrogate endpoint for treatment effect on mortality and hospitalization. However, we found the effect of exercise-based CR on EC to have moderate validity as a surrogate endpoint for treatment effect on HRQOL. Further research is needed to determine whether our findings are generalizable across other HF interventions. Given that the participant-level fitness response to exercise-based CR can be highly heterogeneous (28,29), our findings also need confirmation from individual participant data meta-analyses of exercise-based CR (5).

ADDRESS FOR CORRESPONDENCE: Dr. Oriana Ciani, Institute of Health Research, University of Exeter Medical School, South Cloisters, St. Luke’s Campus, Heavitree Road, Exeter EX1 2LU, United Kingdom. E-mail: o.ciani@exeter.ac.uk.
**COMPETENCY IN MEDICAL KNOWLEDGE:** Using data from randomized trials of exercise-based CR for HF, we formally evaluated the evidence for EC as a surrogate endpoint for mortality, hospitalization, and HRQOL. EC is a poor surrogate endpoint for mortality and hospitalization but has moderate validity as a surrogate for HRQOL. In severe and advanced stages of HF, improvements in independence and the ability to perform daily tasks become even more important than improvements in morbidity and mortality. In this respect, linking increases in EC to HRQOL improvements can be seen as a clinically relevant finding.

**TRANSLATIONAL OUTLOOK:** The use of surrogate endpoints can improve the efficiency of clinical trials by reducing sample size and duration, thereby promoting quicker patient access to new therapies. However, to ensure reliable prediction of final patient-related outcomes, it is important that surrogate endpoints be validated. Our results show an increase of 5 ml/kg/min in VO2peak and 80 m for 6MWT can significantly predict an improvement in HRQOL. Further research is needed to determine whether our findings can be replicated for other types of interventions in chronic HF patients and in other cardiovascular disease populations.

**REFERENCES**

1. McMurray JJ, Adamopoulos S, Anker SD, et al. ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure 2012: the Task Force for the Diagnosis and Treatment of Acute and Chronic Heart Failure 2012 of the European Society of Cardiology. Developed in collaboration with the Heart Failure Association (HFA) of the ESC. Eur Heart J 2012;33:1787-847.


KEY WORDS 6-min walk test, exercise capacity, heart failure, maximum oxygen uptake, surrogate outcomes

APPENDIX For an expanded References section as well as supplemental tables and figures, please see the online version of this paper.