Header: Bootstrapping Exercise Results

Bootstrapping Results of Exercise Therapy and Education for Patients with Congestive Heart Failure

E. Lea Witta
University of Central Florida

Craig Brubaker
Florida Hospital Fish Memorial

Paper presented at the annual meeting of the American Evaluation Association, Reno, November, 2003.

Abstract

When studies are conducted over a period of time, the sample size typically decreases. In a study of the effects of exercise therapy and education with recovering congestive heart failure (CHF) patients[1] (Brubaker, Witta, & Angelopoulus, 2003), the sample size decreased from over 40 to 9 participants after an 18-month time span. Although the quality of life (QOL) scales and measures of physical ability showed improvement, these results were limited to 9 participants. Thus, there was some question concerning the generalizability of this study to other patients with congestive heart failure. The purpose of this study was two-fold: (1) to use the bootstrapping procedure to provide the confidence intervals and estimate generalization error of both physical and quality of life results for that study, and (2) to estimate the probability of sample results as extreme as the observed sample using permutation methods. Most of the findings from this study support the findings from the original study and provide further evidence of the importance of the exercise and education treatment used in the original study. Further, this is not the only study in which improvement in physical measures and QOL constructs have shown improvement. Although the majority of studies on patients with CHF are less than six-months in duration, evidence has been provided that exercise tolerance and QOL can increase significantly following a supervised exercise program (Holst, Kaye, Richardson, Krum, Prior, Aggarwal, Wolfe, & Bergin, 2001; Quittan, Sturm,Wiesinger, Pacher, & Fialka-Moser, 1999; Tokmakova, Dobreva, & Kostianev, 1999). .

Approximately 4.7 million Americans suffer from congestive heart failure (CHF). About seven percent of the total cost of heart disease each year is due to this condition[5] (Heart and Stroke Facts, 2002). In addition, median survival rates following onset of CHF are just 1.7 and 3.2 years for men and women, respectively[6] (Ho, Anderson, Kannel, Grossman, & Levy, 1993). Frequent hospital admissions, poor physical condition, and emotional distress produce a decreased quality of life (QOL) among heart failure patients[7,8] (Cafagna, Ponte, & Burri , 1997; Steptoe, Mohabir, Mahon, & McKenna, 2000). Yet, when studies are conducted over a period of time, the sample size typically decreases. In a study of the effects of exercise therapy and education with recovering congestive heart failure patients[9] (Brubaker, Witta, & Angelopoulos, 2003), the sample size decreased from over 40 to 9 participants after an 18 month time span. Although the quality of life scales and measures of physical ability showed improvement in this study, these results were limited to 9 participants. Thus there was some question concerning the generalizability of this study to other congestive heart failure patients.

The purpose of this study was two-fold: (1) to use the bootstrapping procedure to provide the confidence intervals and estimate generalization error of both physical and quality of life results for that study, and (2) to estimate the probability of sample results as extreme as the observed sample using permutation methods.

Related Literature

One goal of statistical significance testing is to generalize to the population from sample statistics[10] (Mooney & Duval, 1993). “In most cases theory is forced to use asymptotic algebra, producing results that apply only when the sample size is very large”[11] (Kennedy, 1998). If asymptotic results are relied upon when using a small sample, the level of accuracy may not be assured and assumptions pre-requisite to the statistic used may be violated. Thus results produced may be questionable.

The most commonly used method of estimating generalization error is to divide the sample in half, using one half of the data as the “test” set and the other half to estimate the model. One problem with this method is reduction in sample size (see Weiss & Kulikowski, 1991) which, of course, affects statistical significance testing. More recently, some of the statistical software packages (SPSS 11.5) will remove one participant and estimate the model, iteratively until the model has been estimated with each participant removed in a procedure similar to the jackknife method. Although this procedure preserves sample size and is excellent to test the influence of each individual participant, because only one participant is removed each time, it is a poor method of providing generalization error. Monte Carlo methods have also been used to test generalization error. However, when these methods are used, the error estimate is typically drawn from a normal distribution. Consequently the value of these studies in testing generalizability is questionable**.**

Problems in using current methods have lead to increased use of resampling methods. The bootstrap and permutation (randomization) tests are part of these resampling methods. Resampling methods require fewer assumptions than traditional methods and, in most instances are more accurate than classical methods. These methods permit accurate calculation of confidence intervals, and do not require normal distributions or large sample sizes (Howell, 2003). Further, resampling methods are intuitive for each procedure and do not require special new formulas.

Bootstrapping

Bootstrapping is a method of estimating generalization error based on resampling the data rather than using split samples as in cross-validation. In fact, Efron (1983) contended that bootstrapping seemed to work better than cross-validation. In the simplest form of bootstrapping, instead of repeatedly analyzing subsets of the data, you repeatedly analyze sub-samples of the data. Each sub-sample is a random sample with replacement from the full sample (Efron & Tibshirani, 1993).

Bootstrapping permits building a sampling distribution for a statistic by resampling the data used (Fox, 2002). Thus it provides a method of estimating using known data to establish estimates. Further, bootstrapping test statistics provides a sampling distribution tailored to problem. The sampling distribution of a statistic may be estimated empirically without making assumptions about the form of the population. It is a “broadly applicable, nonparametric approach to statistical inference …can be used to derive accurate errors …” (Fox, 1997, p 494). According to Howell (2003) bootstrapping pays attention to population – but has no normality assumption. Thus the data provides a perfect replication of the population and provides a means to create confidence limits on statistics. It can be used to estimate parameters we don’t know how to estimate (ie confidence intervals for a median) or to deal with non-normal distributions. Thus the variability of repeated samples is estimated by drawing repeated samples and observing the distribution.

Permutation (Randomization)

The randomization (permutation) re-sampling procedure answers questions concerning how likely it is that chance sampling error might produce a sample result as extreme as the observed sample. Thus it sets the test condition to the null hypothesis. This provides an exact probability under null hypothesis conditions. It does not rely on assumptions about how the data is distributed and provides accurate results even if the data is skewed (Howell, 2003).

To execute the permutation procedure, conditions for the null hypothesis must be established. If the data consists of two independent groups, scores are randomly selected and assigned to either group. Thus either group may be assigned a score. Consequently, the only variability between groups is chance or random assignment. In repeated measures, scores are time dependent. To establish the null condition, scores remain with the respondent but are permitted to be randomly assigned to a time slot. Thus the null condition of no change over time is established. In either of these cases, the test statistic is then calculated and recorded and the process repeated a given number of times. These results are then used to determine the exact chance probability level of a statistic as large as the one calculated in the original study (Hesterberg, Monaghan, Moore, Clipson, & Epstein, 2000). Although permutation has been deemed the “gold standard” in re-sampling, Westfall and Young (1992) caution that the bootstrap procedure is more powerful.

Method

Data produced by investigating the effects of exercise therapy and education with recovering congestive heart failure patients (Brubaker, Witta, & Angelopoulos, 2003) were used as input for this analysis. Data for each of the nine participants in the study was recorded six times (at entry and at 3 month intervals) on two physical measures (exercise tolerance and MET level) and using the SF-36 Health Survey® which assessed their physical and mental quality of life constructs. Physical constructs consisted of physical functioning, role-functioning physical, bodily pain, and general health. Mental constructs consisted of role-functioning emotional, social functioning, vitality, mental health, and health transition. Data were analyzed using a repeated measures analysis of variance. However, because there were only 9 participants we were concerned that asymptotic results may not apply. Consequently, we reanalyzed this data using bootstrapping and randomization (permutation) tests.

The bootstrapping procedure in Resampling Stats Add-In for Excell was used to produce confidence intervals and to estimate the generalization error. Basically, bootstrapping involves randomly sampling with replacement from the existing data. Replacement permits a single case to be selected ‘n’ times for a sample. Thus, a single case could be randomly selected nine times and used in the analysis. In the current study because the sample size consisted of only nine participants, this means that our sub-sample could consist of one participant listed nine times. Each measure was analyzed independently (quality of life constructs as well as physical measures). As each sub-sample was selected, the scores were submitted to a repeated measures analysis, the resulting F statistic recorded, and a new sub-sample selected. This sampling procedure and analysis was repeated 1000 times for each measure. This entire procedure was repeated a second time.

A permutation test in Resampling Stats Add-In for Excell was used to estimate the probability of sample results as extreme as the observed sample occurring by chance. Permutation tests sample from the existing data without replacement. For this analysis the scores for an individual case remained within that case. However, placement in time for the score could vary. For example, the score recorded for exercise tolerance for a participant during the 18 month (6th time period) trial could be selected as the score used at entry (initial recorded data) or at any of the other time periods. As each sub-sample (participant with scores randomly ordered) was selected, the scores were again submitted to a repeated measures analysis, the resulting F statistic recorded, and a new sub-sample selected. This sampling procedure and analysis was also repeated 1000 times for each measure. This entire procedure was also repeated a second time.

Results and Discussion

Bootstrapping

As is displayed in Table 1, the physical measures used in the Brubaker, Witta, and Angelopoulos (2003) study resulted in similar conclusions when bootstrapping. In bootstrapping analyses, exercise tolerance and MET level never showed a non-significant change (critical value of $F_{5,40}$ =2.45, Kirk, 1995, p. 804). The 95% confidence interval of the F statistic ranged from 13.79 to 49.25 for exercise tolerance and from 29.89 to 77.80 for MET level.

Insert Table 1 about here

There was, however, some discrepancy between conclusions concerning the physical and mental constructs from the QOL instrument. Although three of the physical construct (physical functioning, role functioning physical, and bodily pain) results were similar to the original study ($p \leq 0.05$), the general health construct produced probabilities of 0.073 and 0.087 in the two bootstrapping analyses. The 95% confidence interval for general health ranged from 1.93 to 11.70. Consequently, this construct may not reflect a statistically significant improvement.

In addition, three of the mental constructs (health transition, mental health, and role functioning emotional) were similar to the original study. However, the social functioning construct ($p<0.01$ in the original study) produced a probability level of 0.146 and 0.137 and the vitality construct ($p<0.05$ in the original study) produced a probability level of 0.124 and 0.142 in the bootstrapping procedure. The 95% confidence intervals for social functioning and vitality ranged from 1.44 to 17.24 and 1.50 to 11.23 respectively. These results suggest that three constructs (general health, social functioning, and vitality) from the QOL instrument may not reflect statistically significant improvements as previously thought.

Permutation (Randomization)

In this part of the analysis we were investigating the likelihood of a result as large as the one obtained in the original study occurring by chance. For mental health which was not statistically significant in the original study, the probability of a result as large as the one in the original study by chance was relatively high (71, and 52 of 1000) as is displayed in Table 2. For all other measures the probability of an F statistic as large as the original study by chance was always very small with the highest number of times the original result would be equaled or exceeded in the permutation procedure being 8 of 1000 times ($p<0.01$).

Insert Table 2 about here

Combining the results obtained in the bootstrap and permutation procedures provides a more accurate picture of the findings. Statistically significant changes were not detected in the QOL's mental health construct in the original sample. Mental health was not statistically significant about 40% of the time in the bootstrap procedure and produced results as large as those from the original sample in 7% and 5% of the trials in the permutation procedure. These results are in agreement with findings from the original study and provide further evidence that these measures were not changed significantly.

In addition, when investigating the physical measures (exercise tolerance and MET level) with the bootstrap procedure statistically significant results were always detected. Using the permutation tests, a result as large as the original sample by chance for these measures was never detected. Thus, these results support the statistically significant improvement detected in the original study.

In the original sample results from the QOL's four physical constructs were statistically significant ($p \leq 0.05$). When tested by the bootstrap procedure statistically significant results ($p \leq 0.05$) were detected for bodily pain, physical functioning, and role functioning physical. Results as large as those of the original sample were never detected for role functioning physical or physical functioning and were detected only five and six times of 1000 for bodily pain in the permutation procedure. Although the general health construct produced questionable results in the bootstrap procedure (ie., statistically significant results were not detected in 73 and 87 trials of 1000), when time results were randomly ordered, a result as large as the original sample was only detected in eight of 1000 trials. Consequently, it was concluded that the general health construct was significantly changed as well as the other physical constructs.

Within the mental constructs of the QOL instrument, the bootstrap procedure provided additional evidence that role functioning emotional and health transition were significantly changed by the exercise/education treatment. When tested by randomly ordering the time measures, the F statistic produced by results from these constructs never exceeded the F statistic from the original sample. However, there were some discrepancies within the social functioning and vitality constructs. When re-sampled with replacement, the social functioning results were not statistically significant 146 and 137 times of 1000 trials ($p=.147$, $p=.137$) and the vitality construct produced non-significant results 124 and 142 times of 1000 trials ($p=.124$, $p=.142$). However, when the time measures were randomly ordered, an F statistic as large as the original sample was detected a maximum of 6 times from 1000 trials. Consequently, these results are mixed. The improvement in the health transition and role functioning emotional constructs was statistically significant. However, the social functioning and vitality constructs were still questionable.

Most of these findings support the findings from the original study and provide further evidence of the importance of the exercise and education treatment used in the original study. Further, this is not the only study in which improvement in physical measures and QOL constructs have shown improvement. Although the majority of studies on patients with CHF is less than six-months in duration, evidence has been provided that exercise tolerance and QOL can increase significantly following a supervised exercise program (Holst, Kaye, Richardson, Krum, Prior, Aggarwal, Wolfe, & Bergin, 2001; Quittan, Sturm,Wiesinger, Pacher, & Fialka-Moser, 1999; Tokmakova, Dobreva, & Kostianev, 1999). Our findings support these results. On the other hand, although other long-term studies have found increases in exercise capacity (Gottlieb, Fisher, Freudenberger, Robinson, Zietowski, Alves, Krichten, Vaitkevicus, & McCarter, 1999), improvements in the QOL constructs have not been documented. Our findings have also documented improvement in many of the QOL constructs.

Most major insurance companies do not reimburse cardiac rehabilitation for patients with CHF despite their high hospital readmission rate. However, programs providing a multidisciplinary educational approach have shown to increase the QOL (Afzal, Brawner, & Keteyian, 1998; Brubaker, Witta, & Angelopoulos, 2003) of individuals with CHF, while also improving physical functioning (Koch, Douard, Broustet , 1997; Meyer, Schwaibold, Westbrook, et al, 1997; Tokmakova, Dobreva, Kostianev , 1999). Despite these findings, it is unclear why major insurance companies do not reimburse cardiac rehabilitation programs for CHF patients independent of a current qualifying diagnosis.

References

Afzal, A., Brawner, C.A., & Keteyian, S.J. (1998). Exercise training in Heart Failure. *Progressive Cardiovascular Disease, 41*(3),175-90.

Brubaker, C., Witta, E.L., & Angelopoulos, T.J. (2003). Maintaining exercise tolerance and quality of life by long-term participation in a hospital-based wellness program for individuals with congestive heart failure. *Journal of Cardiopulmonary Rehabilitation, 23*(5), 352-356.

Cafagna, D., Ponte, E., & Burri, R. (1997). The concept of quality of life in cardiac failure. *Minerva Medicine 88*(4), 151-62.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics, 7*, 1-26.

Efron, B. & Tibshirani, R.J. (1993). *Introduction to the Bootstrap*. New York: Chapman & Hall.

Fox, J. (1997). *Applied Regression Analysis, Linear Models, and Related Methods*, Chapter 15. Thousand Oaks: Sage.

Fox, J. (2002). Bootstrapping regression models: Appendix/Companion to *Applied Regression*. Thousand Oaks: Sage.

Gottlieb, S., Fisher, M., Freudenberger, R., Robinson, S., Zietowski, G., Alves, L., Krichten, C., Vaitkevicus, P., & McCarter R. (2001). Effects of exercise training on peak performance and quality of life in congestive heart failure patients. *Journal of Cardiac Failure, 5*(3), 188-94.

Heart and Stroke Facts. (2002). *Statistical Supplement*. American Heart Association. Dallas, TX.

Hesterberg, T. Monaghan, S., Moore, D.S., Clipson, A., & Epstein, R. (2002). *The Practice of Business Statistics*, Companion Chapter 18.

Ho, K., Anderson, K., Kannel, W., Grossman, W., & Levy, D. (1993). Survival after the onset of congestive heart failure in Framingham Heart Study subjects. *Circulation 88*(1).

Holst, D., Kaye, D., Richardson, M., Krum, H., Prior, D., Aggarwal, A., Wolfe, R., & Bergin, P. (2001). Improved outcomes from a comprehensive management system for heart failure. *European Journal of Heart Failure 3*(5), 619-25.

Howell, D. (2003). *Resampling*. [Available www.uvm.edu/~dhowell/StatPages /Resampling /Resampling.html] (Includes randomization and bootstrapping).

Kennedy, P.E. (1998). Bootstrapping student understanding of what is going on in econometrics. *Journal of Economic Education*, 110-122.

Kirk, (1995). *Experimental Design: Procedures for the behavioral sciences*. Pacific Grove: Brooks/Cole Publishing. p 804.

Koch, M., Douard, H., & Broustet, J.P. (1997). The benefit of graded physical exercise in chronic heart failure. *Chest.101*(5S), 231S-235S.

Meyer, K., Schwaibold, M., Westbrook, S., et al. (1997). Effects of exercise training and activity restriction on 6-minute walking test performance in patients with chronic heart failure. *American Heart Journal, 133*(4), 447-53.

Mooney, C.Z., & Duval, R.D. (1993). *Bootstrapping: A nonparametric approach to statistical inference*. Sage Quantitative Applications in the Social Science Series, 95. Thousand Oaks: Sage.

Quittan, M., Sturm, G., Wiesinger, B.F., Pacher, R., & Fialka-Moser, V. (1999). Quality of life in patients with chronic heart failure: a randomized controlled trial of changes induced by a regular exercise program. *Scandinavian Journal of Rehabilitative Medicine, 31*(4), 223-8.

Steptoe, A., Mohabir, A., Mahon, N.G., & McKenna, W.J. (2000). Health related quality of life and psychological wellbeing in patients with dilated cardiomyopathy. *Heart, 83*(6), 645-50.

Tokmakova, M., Dobreva, B., & Kostianev, S. (1999). Effects of short-term exercise training in patients with heart failure. *Folia Medicine, 41*(1), 68-71.

Table 1

Results of Bootstrap Procedures using the Health Data

Measure	$F_{observed}$	CI_{lower}	CI_{upper}	Bootstrap1[a]	Bootstrap2[a]
Physical Measures					
Exercise Tolerance	20.21**	13.79	49.25	0	0
MET Level	39.8**	29.89	77.80	0	0
Physical Constructs					
Bodily Pain	4.23**	2.10	11.62	45	42
General Health	3.74**	1.93	11.70	73	87
Physical Functioning	13.09**	8.56	30.13	0	0
Role Functioning Physical	7.03**	3.42	20.61	7	4
Mental Constructs					
Health Transition	9.85**	5.35	26.97	0	0
Mental Health	2.09	0.85	6.96	388	429
Role Functioning Emotional	8.18**	3.70	26.62	6	4
Social Functioning	4.45**	1.44	17.24	146	137
Vitality	3.43*	1.50	11.23	124	142

Note. $F_{observed}$ F statistic from the original sample. CI_{lower} Bootstrap lower limit of the 95% confidence interval. CI_{upper} Bootstrap upper limit of the 95% confidence interval. [a]Number of times in 1000 trials the results were NOT statistically significant when respondents were randomly sampled with replacement. * $p \leq .05$. ** $p \leq .01$.

Table 2

Results of Permutation Procedures on the Health Data

Measure	$F_{observed}$	Permutation 1[a]	Permutation 2[a]
Physical Measures			
Exercise Tolerance	20.21**	0	0
MET Level	39.8**	0	0
Physical Constructs			
Bodily Pain	4.23**	5	6
General Health	3.74**	8	8
Physical Functioning	13.09**	0	0
Role Functioning Physical	7.03**	0	0
Mental Constructs			
Health Transition	9.85**	0	0
Mental Health	2.09	71	52
Role Functioning Emotional	8.18**	0	0
Social Functioning	4.45**	6	1
Vitality	3.43*	6	2

Note. $F_{observed}$ F produced by repeated measures analysis of the original sample. [a]Number of times in 1000 trials the results were as large as the original sample by chance when order of the data for each respondent was randomly shuffled. * $p \leq .05$. **$p \leq .01$.