Published in final edited form as:

Med Sci Sports Exerc. 2003 May ; 35(5): 736–740. doi:10.1249/01.MSS.0000064995.89335.40.

The Illusion of Improved Physical Fitness and Reduced Mortality

Paul T. Williams

Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720

Paul T. Williams: PTWilliams@LBL.gov

Abstract

Purpose—To conduct a computer simulation to assess the effects of measurement error on prospective epidemiological studies that attribute mortality outcomes to apparent changes in the independent variable (e.g., physical fitness or activity) at baseline.

Methods—As an example, we evaluated the design of the Aerobics Center Longitudinal Study (ACLS). This study compared apparent changes in fitness between two baseline visits to mortality during a subsequent 5-yr follow-up period. Unfit men who were reclassified as fit at the second baseline examination (6.6% of sample) and fit men who were reclassified as unfit (2.3%) had follow-up mortality rates that were between those of men who were consistently classified as fit or unfit. This study design was simulated assuming that differences between baseline treadmill test durations were due to measurement error alone. Based on our own data, we estimated that repeat measurements of treadmill test duration have correlation of $r = 0.89$ in the absence of any real fitness change.

Results—There is excellent agreement between the published ACLS risk reductions and our simulated reductions for both cardiovascular disease (CVD) and total mortality. Compared with the “Unfit→Unfit” (the referent group), the estimated relative risks from the simulations for men who were reclassified as fit (i.e., “Unfit→Fit”) were 0.57 for total mortality and 0.52 for CVD mortality, and for men who remained classified as fit (“Fit→Fit”), they were 0.33 for total mortality and 0.20 for CVD mortality.

Conclusion—The imprecision of the fitness measurement alone (i.e., measurement error) is sufficient to produce the reported ACLS risk reductions in initially unfit men who get reclassified as fit in a subsequent clinic visit. This statistical artifact will apply to other studies that use this design.

Introduction

To identify causes of disease, a standard epidemiological study design is to measure a set of variables at baseline and to compare them with health events (e.g., death, disease, and illness) during a period of follow-up. For example, in their landmark article on physical fitness and mortality, Blair et al. [1] showed that men and women who had low cardiorespiratory fitness at baseline were more likely to die during a follow-up period than those having greater fitness.

One modification of this approach is to measure the variables at two separate occasions during the baseline period. The apparent change in the variables between the first and second baseline visits is then compared with health outcomes during the follow-up period. Results from this second design are used to infer that risk factor changes are related to changes in health outcomes. Because change is involved, the findings are interpreted as providing stronger evidence of causality than findings from a single measurement. Studies that have employed this design, or a variation thereof, include studies of Norwegian men [4], older men from the British Regional Heart Study [14], the Harvard Alumni Study [7], women from Gothenburg Sweden [6], men and women from the Corpus Christi Heart Project [12], and the Aerobics Center Longitudinal Study (ACLS) [2]. For example, the ACLS measured physical fitness at two separate baseline visits that were 4.9 yr apart [2]. The apparent changes in fitness between

the two baseline visits were then compared with mortality during a subsequent 5.1-yr follow-up period. Specifically, unfit men who were reclassified as fit at the second baseline examination (6.6% of sample) and fit men who were reclassified as unfit (2.3% of sample) had follow-up mortality rates that were between those of men who were consistently classified as fit or unfit (Fig. 1A). This led the authors to conclude that "... patients will reduce risk of mortality by increasing physical activity and improving fitness."

In this report, we propose an alternative explanation that attributes the ACLS observations to statistical artifact due to measurement error rather than any real association between improved fitness and mortality. Specifically, we view the study as a traditional prospective design with two baseline measurements (Fig. 1B), rather than a study of changing fitness and risk (Fig. 1A). Our analyses show that measurement error alone is sufficient to explain the intermediate risks for unfit individuals who are reclassified as fit, and fit individuals who are reclassified as unfit. This type of statistical artifact (a variation on regression to the mean) is germane to all epidemiological studies that employ this design.

Methods

For clarity, we focus on a specific study to illustrate the model. The Aerobic Center Longitudinal Study measured cardiorespiratory fitness by the length of time a participant was able to remain on a motorized treadmill (treadmill test duration). In the present report, we employ a computer simulation of treadmill test durations to evaluate how measurement error could impact the ACLS findings. For each simulated individual "i", we created two treadmill test durations (O_{i1} and O_{i2}) that corresponded to the two ACLS baseline observations. The first observed treadmill duration (O_{i1}) was the sum of two random normal numbers: the first representing an individual's true fitness level (T_i) and the second representing the measurement error (E_{i1}) which when added to their true fitness gave the observed duration ($O_{i1} = T_i + E_{i1}$). To create the second treadmill duration (O_{i2}) a second modicum of error (E_{i2}) was added to the individual's true fitness level (T_i); i.e., $O_{i2} = T_i + E_{i2}$. Note that the underlying true fitness level (T_i) did not change between surveys but that the observed durations at survey 1 (O_{i1}) and survey 2 (O_{i2}) did because they have different random errors (E_{i1} and E_{i2} , respectively). Our own unpublished data in sedentary men suggests that treadmill test durations are normally distributed (i.e., in 81 men [17], the skewness was 0.04 and kurtosis was 2.64, see [11]). In our data, the treadmill test durations taken 1 yr apart had a correlation of 0.89 [16]. Therefore, by setting the variance of the true fitness scores (T_i) to 0.89 and the variances of the measurement errors to 0.11, we were able to simulate pairs of normally distributed random numbers (O_{i1} and O_{i2}) that have a mean of "0," a standard deviation of "1" and a correlation of " $r = 0.89$." These numbers were rescaled to have a mean of 16.9 and a standard deviation of 4.7 min to correspond to the distribution of actual treadmill test durations (see Fig. 2 top; the scaling is immaterial and has no affect on the analyses to follow). The simulated durations that fell below and above the 20th percentile were classified as unfit and fit, respectively, to match what was done in the ACLS. One million pairs of observation were generated.

The next step was to ascribe a disease risk to each simulated treadmill duration. This was done by referring to the seminal paper relating mortality to the first baseline duration reported by Blair et al. [1]. In that article, the authors reported that the first through fifth quintiles of fitness had age-adjusted total mortality rates that were 64.0, 25.5, 27.1, 21.7, and 18.6 deaths per 10,000 man-years. The age-adjusted cardiovascular disease (CVD) mortality rates per 10,000 man-years were 24.6 for the least fit quintile, 7.8 for the second and third quintiles, and 3.1 for the fourth and fifth quintiles (the most fit). In our simulation, we postulated that the risk depended upon the true underlying fitness score and not upon the random error. The first through fifth quintiles of the baseline simulated treadmill durations (O_{i1}) had mean true values of 10.70, 14.54, 16.90, 19.27, and 23.11 min (i.e., these were the average true fitness scores

for the first, second, third, fourth, and fifth quintiles of the simulated durations). A risk function was calculated using line segments that connected the reported risks and the mean true fitness values for each quintile of observed treadmill duration (Fig. 3). This function was used to calculate the total mortality and CVD mortality risks per 10,000 man-years from each individual's true fitness score. The event rates for the 1989 follow-up survey [1] (and correspondingly our simulation) are about half that reported for the ACLS change cohort [2], reflecting presumably the age difference of the samples; however, this is inconsequential because the analyses are based on relative risk.

Results

One million observations were simulated assuming that there was no change in true fitness but that random error alone produces differences between the first and second treadmill test duration measurements. Figure 2 (top) shows a subset of 10,000 points. Each plotted point represents a simulated treadmill duration at survey 1 (O_{i1}) plotted along the x-axis and at survey 2 (O_{i2}) plotted along the y-axis. The solid line along the diagonal represents the loci of true fitness scores (T_i), which in our model remain constant within each individual. The figure presents a cloud of points flanking the diagonal as would be expected from repeated measurements. The upper right (Fit→Fit) and lower left quadrants (Unfit→Unfit) include individuals who did not change their fitness classification. The shaded regions represent those individuals who went from being classified as unfit to fit (upper left quadrant) or fit to unfit (lower right quadrant).

Figure 2 bottom shows that the simulation yielded the following percentage composition of men: 14.7% consistently unfit (Unfit→Unfit), 10.5% apparent changers (5.25% each in the Unfit→Fit and Fit→Unfit groups), and 74.8% consistently fit (Fit→Fit) with average true treadmill time fitness scores (T) of 10.2, 13.3, and 18.7 min, respectively.

Figure 4 compares the ACLS published relative risks for individuals who purportedly maintained and changed their fitness categories (solid bars) [2] and the expected relative risks from our simulations (hatched bars), i.e., assuming that these individuals had not actually changed their fitness but who were occasionally reclassified due to random measurement error. There is excellent agreement between the published ACLS risk reductions and our simulated reductions for both CVD and total mortality. We estimated that total mortality per 10,000 man-years was 72.61 for those classified as unfit on both occasions, 41.21 for fit individuals who became unfit (the same for unfit who became fit) and 23.78 for fit individuals who remained fit. For CVD mortality, the rates per 10,000 man-years were 28.36 for consistently unfit individuals, 14.62 for individuals who changed their fitness category, and 5.67 for those who were classified as consistently fit. Compared with the “Unfit→Unfit” (the referent group), the estimated relative risks for men who were reclassified as fit (i.e., “Unfit→Fit”) were 0.57 for total mortality and 0.52 for CVD mortality, and for men who remained classified as fit (“Fit→Fit”) they were 0.33 for total mortality and 0.20 for CVD mortality.

The above simulation assumed a correlation of $r = 0.89$ between treadmill test durations measured at two separate occasions. We changed the simulations to assess the sensitivity of the results to this assumption. Decreasing the test-retest correlation from 0.89 to 0.70 had little effect, i.e., the relative risks for “Unfit→Fit” were 0.57 for total mortality and 0.52 for CVD mortality, and for “Fit→Fit” they were 0.29 for total mortality and 0.21 for CVD mortality. Increasing the test-retest correlation from 0.89 to 0.95 also had little effect, i.e., the relative risks for “Unfit→Fit” were 0.57 for total mortality and 0.52 for CVD mortality, and for “Fit→Fit” they were 0.34 for total mortality and 0.21 for CVD mortality.

Discussion

Prospective epidemiological studies that compare apparent changes in the independent variables at baseline to subsequent mortality are potentially very influential. Because change is involved, the findings are interpreted as providing stronger evidence of causality than findings from a single measurement. For example, the ACLS paper is among the most widely cited epidemiological studies of physical activity and health. In the 3 yr from 1998 to 2000, it ranked second (148 citations) among the 96 cohort, population-based, and epidemiological studies listed in Chapter 4 of the 1996 Surgeon General's Report on Physical Activity [13] (only Blair's et al. initial 1989 report on cardiorespiratory fitness and mortality with 209 citations was cited more frequently) [10].

The ACLS wide acclaim reflects in part two profound implications for public health policy. First, because the ACLS specifically involved men who appeared to move in or out of the lowest quintile of fitness, the results would suggest that increasing fitness among the least fit individuals (presumably through modest increases physical activity) produces the greatest health benefit. Second, the results would suggest that the high mortality among low fit men is not due to a genetic predisposition as improved fitness lowered mortality and worsened fitness raised mortality, in the same individual.

We believe that a flaw in the interpretation of study designs such as that employed by the ACLS is the failure to recognize that the independent variables (e.g., physical fitness) are measured with error. A scatterplot for two normally distributed random variables having a correlation of 0.89 is presented in Figure 2. The scatterplot of points about the diagonal would be expected for any biological measurement with less than perfect measurement. The plot may be divided into quadrants at the 20th percentiles of the x (horizontal) and y (vertical) axis. This corresponds to the division between the fit and unfit as defined by the ACLS. The ACLS authors attributed the upper left and lower right quadrants as having increased or decreased fitness. However, unless the first and second measurements were perfectly correlated, the upper left quadrant and lower right quadrants must contain observations. It is not necessary for these to arise for any other reason than measurement error (Fig. 2). There is good agreement between the number of observations expected to fall within these quadrants from measurement error alone and the numbers presented by the ACLS authors (8.9% in ACLS vs 10.5% in our simulations, although the ACLS observations were not symmetrically distributed in the two quadrants).

In our simulations, we estimated the probability for death from the tables published from the 1989 prospective study on mortality and cardiorespiratory fitness [1]. The simple approach assigns the probability of death on the true mean fitness score which remained constant over time (Fig. 3). We then calculated the expected risk for the four quadrants of Figure 2 (top). The true fitness scores for the upper left (Unfit→Fit) and lower right quadrants (Fit→Unfit) were intermediate to the true fitness scores of the lower left (Unfit→Unfit) and upper right (Fit→Fit) quadrant (Fig. 2 bottom). As depicted in the Figure, men who were reclassified are actually those who were on or near the boundary between fit and unfit, and were likely to be reclassified each time they are measured. Because their true fitness was intermediate between the always fit and always unfit, their mortality risk was also intermediate (Fig. 4). Similar results were obtained if the correlation between treadmill durations was increased to 0.95 or decreased to 0.80. One simple interpretation of our simulation is that the average of two baseline treadmill durations more accurately assigns an individual's fitness than a single duration [16].

Measurement error is a phenomenon that is known to require consideration before any other attributions [9]. The effects reported here occur for the same reason that regression to the mean occurs: observations that are at the extremities arise because their true value is actually extreme, they have an extreme measurement error, or some of both. Subsequent measurements on the

same individual will continue to be extreme only if the true value was extreme, but not if the measurement error was extreme (the measurement errors being independent of one another). Thus, treadmill durations that remain low on both measurements probably represent truly unfit individuals that have high mortality risk, and those that were low on the first measurement and higher on the second measurement were probably initially classified as unfit in part because of measurement error, and their true fitness will be higher than that of the consistently unfit group. However, the true fitness (and mortality) of those that changed fitness categories will be less than those who consistently exhibit high treadmill duration scores.

In conclusion, the effects of measurement error we describe in this report are not obvious and the authors of the ACLS report cannot be faulted for not anticipating their effect. This example demonstrates the difficulty of inferring change from nonexperimental data. There is clear evidence that fitness is related prospectively to mortality rates [1,8]. However, the claim that changes in fitness alter rates of mortality requires further justification [2]. This is important because the purported association between increased (decreased) physical fitness and reduced (increased) mortality has been cited widely in support of current health policies [5,13]. The argument for favoring physical activity interventions for mainly persons who are the least active relies heavily upon studies of fitness, and the presumption that these studies are germane to the setting of physical activity recommendations. Elsewhere, we have shown that the CVD risks associated with being unfit and being inactive are significantly different from each other [15].

We have demonstrated that the reported changes in risk from moving in or out of the lowest fitness quintile in the ACLS can be attributed entirely to measurement error. Although physical activity increases fitness and may be an appropriate therapy for the unfit, cardiac deficiency or genes rather than inactivity may be the root cause for the very unfit. In sedentary subjects, maximal heritability estimates of 50% are reported for physical fitness, which may be an underestimate since the estimate includes the attenuating effects of measurement error [3]. We believe that it remains to be determined whether targeting the least active individuals for public health interventions will produce the greatest reductions in mortality.

Acknowledgments

This work was supported in part by grant HL-58621 from the National Heart Lung and Blood Institute and was conducted at the Lawrence Berkeley National Laboratory (Department of Energy DEAC03- 76SF00098 to the University of California).

References

1. Blair SN, Kohl HW 3rd, Paffenbarger RS Jr, Clark DG, Cooper KH, Gibbons LW. Physical fitness and all-cause mortality. A prospective study of healthy men and women. *JAMA* 1989 Nov 3;262(17): 2395–401. PubMed PMID: 2795824. [PubMed: 2795824]
2. Blair SN, Kohl HW 3rd, Barlow CE, Paffenbarger RS Jr, Gibbons LW, Macera CA. Changes in physical fitness and all-cause mortality. A prospective study of healthy and unhealthy men. *JAMA* 1995 Apr 12;273(14):1093–8. PubMed PMID: 7707596. [PubMed: 7707596]
3. Bouchard C, Daw EW, Rice T, Pérusse L, Gagnon J, Province MA, Leon AS, Rao DC, Skinner JS, Wilmore JH. Familial resemblance for VO₂max in the sedentary state: the HERITAGE family study. *Med Sci Sports Exerc* 1998 Feb;30(2):252–8. PubMed PMID: 9502354. [PubMed: 9502354]
4. Erikssen G, Liestøl K, Bjørnholt J, Thaulow E, Sandvik L, Erikssen J. Changes in physical fitness and changes in mortality. *Lancet* 1998 Sep 5;352(9130):759–62. PubMed PMID: 9737279. [PubMed: 9737279]
5. Fletcher GF, Balady G, Blair SN, Blumenthal J, Caspersen C, Chaitman B, Epstein S, Sivarajan Froelicher ES, Froelicher VF, Pina IL, Pollock ML. Statement on exercise: benefits and recommendations for physical activity programs for all Americans. A statement for health

professionals by the Committee on Exercise and Cardiac Rehabilitation of the Council on Clinical Cardiology, American Heart Association. *Circulation* 1996 Aug 15;94(4):857–62. PubMed PMID: 8772712. [PubMed: 8772712]

6. Lissner L, Bengtsson C, Björkelund C, Wedel H. Physical activity levels and changes in relation to longevity. A prospective study of Swedish women. *Am J Epidemiol* 1996 Jan 1;143(1):54–62. PubMed PMID: 8533747. [PubMed: 8533747]
7. Paffenbarger RS Jr, Hyde RT, Wing AL, Lee IM, Jung DL, Kampert JB. The association of changes in physical-activity level and other lifestyle characteristics with mortality among men. *N Engl J Med* 1993 Feb 25;328(8):538–45. PubMed PMID: 8426621. [PubMed: 8426621]
8. Physical activity and cardiovascular health. NIH Consens Statement 1995 Dec 18-20;13(3):1–33. Review. PubMed PMID: 9160011.
9. Plewis, I. *Analyzing Change: Measurement and Explanation Using Longitudinal Data*. New York: John Wiley & Sons; 1985. p. 73-88.
10. Science Citation Index. Philadelphia: ISI; January through December. 2000 2001
11. Snedecor, GW.; Cochran, WG. *Statistical Methods*. Vol. 6th. Ames IA: Iowa State University Press; 1967. p. 552
12. Steffen-Batey L, Nichaman MZ, Goff DC Jr, Frankowski RF, Hanis CL, Ramsey DJ, Labarthe DR. Change in level of physical activity and risk of all-cause mortality or reinfarction: The Corpus Christi Heart Project. *Circulation* 2000 Oct 31;102(18):2204–9. PubMed PMID: 11056093. [PubMed: 11056093]
13. U.S Department of Health and Human Services. *Physical activity and health: a report of the Surgeon General*. Atlanta: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion; 1996. p. 81-172.
14. Wannamethee SG, Shaper AG, Walker M. Changes in physical activity, mortality, and incidence of coronary heart disease in older men. *Lancet* 1998 May 30;351(9116):1603–8. PubMed PMID: 9620713. [PubMed: 9620713]
15. Williams PT. Physical fitness and activity as separate heart disease risk factors: a meta-analysis. *Med Sci Sports Exerc* 2001 May;33(5):754–61. PubMed PMID: 11323544. [PubMed: 11323544]
16. Williams PT. Dose of exercise and health benefit. *Arch Intern Med* 1997 Aug 11-25;157(15):1774–1776. PubMed PMID: 9250246. [PubMed: 9250246]
17. Wood PD, Haskell WL, Blair SN, Williams PT, Krauss RM, Lindgren FT, Albers JJ, Ho PH, Farquhar JW. Increased exercise level and plasma lipoprotein concentrations: a one-year, randomized, controlled study in sedentary, middle-aged men. *Metabolism* 1983 Jan;32(1):31–9. PubMed PMID: 6848894. [PubMed: 6848894]

A. Original interpretation (2)

B. Our interpretation

* age-adjusted all-cause death rate per 10,000 man-years

Figure 1.

Two interpretations of the design of the Aerobic Center Longitudinal Study: (A) by the original authors [2] and (B) by this report.

Figure 2.

Top. Scatterplot of 10,000 points having a bivariate normal distribution with a mean of 16.9 and a standard deviation of 4.7 min. In the measurement error model, the true fitness levels lie along the diagonal and the observed treadmill durations measured on the first (x-axis) and second (y-axis) clinic visits are represented by the points. By random error alone, 10.5% of individuals are reclassified into a different fitness category, with no true change in fitness. Bottom. Distribution of true fitness scores in individuals who are consistently fit, consistently unfit, and apparently change their fitness category in the measurement error model. Note that the apparent changers have a distribution of true fitness that is intermediate those classified as consistently fit and consistently unfit.

Figure 3. Interpolating of the risk for total mortality and CVD using line segments linking the mean true fitness and published risk [1] for each quintile.

Figure 4.

Published relative risks from ACLS for men who remained fit, purportedly went from fit to unfit, purportedly went from unfit to fit, and men who remained consistently unfit [2]. There is excellent agreement between the published relative risks and the expected relative risks, assuming that all reclassification is due to measurement error.