REVIEW

Biostatistical Analysis: A Primer for Clinical Exercise Physiology, Part 1

Suzanne L. Havstad, MA[1], George W. Divine, PhD[1]

ABSTRACT

In this first of a two-part series on introductory biostatistics, we briefly describe common designs. The advantages and disadvantages of six design types are highlighted. The randomized clinical trial is the gold standard to which other designs are compared. We present the benefits of randomization and discuss the importance of power and sample size. Sample size and power calculations for any design need to be based on meaningful effects of interest. We give examples of how the effect of interest and the sample size interrelate. We also define concepts helpful to the statistical inference process. When drawing conclusions from a completed study, *P* values, point estimates, and confidence intervals will all assist the researcher. Finally, the issue of multiple comparisons is briefly explored. The second paper in this series will describe basic analytical techniques and discuss some common mistakes in the interpretation of data. *Journal of Clinical Exercise Physiology*. 2018;7(3):63–69.

Keywords: biostatistics, sample size, randomization, *P* value

INTRODUCTION

To a large extent, the amount of useful information that can be extracted from research data is a function of the statistical methods used in analyzing it. The correct analysis of data has its foundation in the design phase of a research project, where proper estimation of sample size is essential. Heath (1) discusses epidemiological considerations in exercise physiology research and describes many common research designs, their strengths, and their limitations.

This paper is the first of a two-part series on introductory biostatistical concepts. This first paper reviews design considerations that are important to sound data analysis, with an emphasis on the role of randomization, power, and sample size calculations. Basic components of statistical inference, including the issue of multiple comparisons, are also presented.

STUDY DESIGN

Biostatisticians provide important help with the design and implementation of a study. Specifically, while the researcher provides the impetus for the entire research process, the biostatistician helps ensure that the question asked has a good chance of being answered, given the proposed study design.

The randomized clinical trial (RCT) is considered the gold standard for medical research. However, cross-sectional, case-control, prospective or retrospective cohort, and intervention studies also play a role in research (1). The study questions or aims often dictate the design of the study or, conversely, due to resource or ethical considerations, the design of the study may dictate which research questions can be answered. Heath (1) has described four of these design types. The key characteristics of all six common study designs (2-4) are highlighted in Table 1.

Briefly, cross-sectional studies are often the easiest to implement, are helpful in obtaining estimates of prevalence rates of disease among different populations, and may be the first step in exploring the disease process. A case-control design generally cannot address questions of incidence or prevalence. Instead, it is commonly used when an association with disease is uncertain or the disease is rare. A prospective cohort study is a longitudinal study in which the researcher observes characteristics of a

[1]Department of Public Health Sciences, Henry Ford Hospital, Detroit, MI 48202, USA

Address for correspondence: George W. Divine, Department of Public Health Science, Henry Ford Hospital, One Ford Place, Detroit, MI 48202 USA; e-mail: gdivine1@hfhs.org.

Conflicts of Interest and Source of Funding: None.

Table 1. Characteristics of common study designs

Design	Retrospective	Prospective	Randomization	Experimental Study	Observational Study	Usually Lower Cost	What Can Be Shown?	Other Attributes of Design
Cross-sectional	√				√	√	Associations	Independent and dependent variables measured *at same time*
Prospective cohort		√			√		Associations	Permits observation of characteristics or behaviors over time
Retrospective cohort	√				√	√	Associations	Efficient for cohort investigation of diseases with long latency periods
Case-control	√				√	√	Associations	Subjects selected based on whether they do (cases) or do not (controls) have a particular disease; good for studying rare diseases
Intervention study		√		√		√	Associations, potential intervention effect	Sometimes called "quasi-experimental"; intervention on one group only
Randomized clinical trial (RCT)		√	√	√			Causation, intervention effect	Sample size and power must be adequate to detect the desired difference between study groups

cohort of people over time. However, since a prospective cohort study is observational rather than experimental, it (like cross-sectional or case-control trials) cannot prove a causal relationship.

A retrospective cohort study differs from the prospective cohort in that the outcome of interest has already occurred. The retrospective cohort study can usually be conducted more quickly and cost-effectively than the prospective counterpart, but it has more disadvantages. In particular, retrospective studies are often dependent on data previously collected for other purposes and, as such, may not include all factors of interest; these studies may be incomplete or have inconsistent information for some of the subjects (4).

Intervention studies or clinical trials without a comparison group are useful as pilot studies or for obtaining descriptive data. Their advantage is that an intervention is given to a set of individuals, thus making it plausible to examine causality. However, due to the many possible biases in this type of study, causality can rarely be concluded with certainty. Hill (5) lists nine criteria for assessing causality that could apply to this type of design. Unfortunately, even Hill concludes that support for all nine may not "bring indisputable evidence for or against the cause-and effect hypothesis" (5). Intervention studies may be an informative and affordable step to take prior to full-scale investment in an RCT.

The missing ingredients from the intervention study design are a control/comparison group, along with the important benefits that randomization provides. With a sufficient sample size, randomization will protect against the bias that can occur if anyone consciously or unconsciously assigns subjects with a better prognosis to the intervention group they feel will be superior. When feasible, proper randomization methods will blind treatment assignment to anyone involved in the selection or allocation process. Nonrandomized trials are always open to the suspicion of investigator or subject selection bias. The other benefit of randomization is that when the sample size is sufficiently large, known and unknown factors that may influence outcome (i.e., possible confounders) will be equally distributed among the study groups. While it may be possible to adjust for known confounders in most design types, only a randomized trial can fully control unknown confounders.

An RCT uses prospective data collection, includes a randomized control group, and can prove causation. Since randomization can ensure that groups are comparable prior to the intervention, effects can be cleanly attributed to the intervention. It is for reasons such as these that the RCT is considered the gold standard of research design (6).

SAMPLE SIZE AND POWER

An important consideration in designing a study is the calculation of the sample size required to have a reasonable chance of statistical significance when there truly is an important difference or association. To effectively discuss sample size calculations, several key terms need to be defined (see Common Statistical Terms box).

The *alternative or research hypothesis* (H_A) is a statement about the difference or association one wants to

COMMON STATISTICAL TERMS

Term	Definition
Null hypothesis (H_0)	Statement about the value of the population parameter (often states there is "no difference")
Alternative or research hypothesis (H_A)	Statement that contradicts null hypothesis (often states there is "some difference")
Type I error	Mistakenly rejecting the null hypothesis (H_0)
Alpha (α)	Probability of making a Type I error
Type II error	Not rejecting H_0 when H_0 is false
Beta (β)	Probability of making a Type II error
Power (1-β)	Probability that a significant difference will be found, given that H_A is true
Two-sided test	Tests for a difference from the null hypothesis that can go in either direction (> or <)
Normal distribution	Theoretical distribution of values that is symmetrical, unimodal, and bell shaped
Variance	Measurement of the "typical" squared distance the values have from the sample mean
Standard deviation	Commonly used measure of the spread or dispersion of data. The square root of the variance is the standard deviation
Standard error	Standard deviation divided by the square root of *n* (sample size)
Dependent variable	Variable that is identified as an effect, result, or outcome variable. Is sometimes viewed as being caused by the independent variable
Independent variable	Variable that is identified as a possible causal variable
Confidence interval	Provides a range of values that are intended to contain the parameter of interest with a specified degree of confidence
P value	Probability of obtaining a result as extreme or more extreme than the actual sample value obtained, given that H_0 is true
Prospective study	Study subjects are followed forward in time
Retrospective study	Subjects' characteristics determined from existing information
Observational study	Observe the characteristics of subjects with no intervention by an investigator
Experimental study	Impose certain characteristics on at least one group

demonstrate. For example, a research hypothesis may be that supervised exercise sessions will improve systolic blood pressure (SBP) in patients with heart failure, compared with a similar group with no supervised exercise sessions provided. The *null hypothesis* (H_0) is a statement of no difference, for example, the supervised exercise group will have the same change in SBP at the end of the study as the standard-care group.

Power is defined as the probability of rejecting the null hypothesis, based on the observed data, when the alternate hypothesis is assumed to be true. In other words, power is the probability that a statistical test will yield statistically significant results, given that the association really does exist. By convention, a study is often powered at 80% or 90% (6,7) to yield a reasonable chance (i.e., 80%–90%) of detecting a difference if it truly exists.

There are two types of errors that can occur when making conclusions from results of a statistical test (Table 2). The Type I error, denoted as alpha (α), rejects the null hypothesis when there truly is no association or difference. The Type II error, denoted as beta (β), is defined as not rejecting the null hypothesis when it is false. Restated, power is the probability of *not* making a Type II error. A power of 80% implies a 20% probability of making a Type II error.

TABLE 2. Possible outcomes of hypothesis testing.

Decision From Statistical Test	Reality	
	H_0 True	**H_0 True**
Fail to reject H_0 (accept H_0)	Correct decision	Type II error (β)
Reject H_0	Type I error (α)	Correct decision

H_0: null hypothesis.

TABLE 3. Sample size calculation example.

Average Decrease in SBP in Control Group	Average Decrease in SBP in Supervised Exercise Group	*N* per group	Clinically Relevant Difference?
0	2	2103	Probably not
0	5	338	Maybe
0	8	133	Likely
0	10	86	Clearly

By convention, α is usually set to be 0.05. Obviously, researchers would like the chance of both Type I and Type II errors to be small. However, holding all other assumptions constant, when one type of error is decreased, the other type is increased. Comparing the typical relative sizes of β (0.20 or 0.10) to α (0.05) reflects the greater seriousness with which a Type I error is regarded.

A study should be powered high enough to have a reasonable chance (80%–90% probability) of confirming the primary hypothesis if it is true. The power depends greatly on the sample size. The exact form of the sample size formula depends on the statistical test planned for the analysis (7). Note that studies can be designed to have power to detect the effects for all secondary hypotheses as well, but this can be difficult and costly.

Table 3 shows an example of sample site calculations for the research question: Does SBP change after supervised exercise sessions in patients with heart failure? For these calculations, it was assumed there would be no change in the control group but that there would be hypothetical average SBP decreases between 2 and 10 in the exercise group. The average decrease was varied for the exercise group to demonstrate the different sample sizes that would be required to have 90% power to detect those differences. As illustrated, the larger the sample size, the smaller the average decrease that has a reasonable chance of being detected.

In Table 3, the statistical test assumed in the sample size calculations is a two-sample Student's *t* test (to be discussed in Part 2 of this series). This statistical test was chosen because the analysis plan states that the mean difference in SBP between the exercise and control groups will be tested with a Student's *t* test. To calculate power, one needs an estimate of the common variability in SBP among such patients. For instance, Keteyian and colleagues (8) reported an estimate of 20 for a standard deviation for SBP among similar patients. Therefore, using an estimated standard deviation of 20, a power of 90%, α level of 0.05, and two-sided testing, sample size calculations were performed that generated the numbers as presented in Table 3.

EFFECT OF INTEREST

The final critical element of a sample size calculation is the minimal clinically important difference, also called the *effect of interest*. An effect of interest can be selected based on input from the literature, from pilot study data, or from the investigator's clinical judgment after years of experience in the field. For example, in the scenario used to generate Table 3, it is of interest to see a useful decrease in SBP in the supervised exercise group when compared to the standard-care group. The question the researcher must consider is, what amount of a decrease in the exercise group is clinically important? Most investigators would agree that a 2-point decrease in systolic blood pressure is not clinically useful, whereas a 10-point decrease is of clinical interest. The gray area is the 5- to 8-point difference.

As Table 3 shows, there is an inverse relationship between the effect of interest and sample size: The smaller the effect, the larger the sample size required, and vice versa. As another example, Table 4 shows a secondary prevention study designed with 90% power of detecting a 10% 5-year reinfarction rate in a supervised exercise group versus a 30% 5-year reinfarction rate in a no-exercise control group. In this example, approximately 82 subjects would be required per group, assuming two-sided testing and an α level of 0.05 (9). The required sample size increases when the difference of interest decreases. It is much harder to detect smaller differences because the element of chance or random variation is so large relative to the intervention effect. However, an overpowered study can find an unimportant small effect when a larger one is really of clinical interest, thus almost certainly wasting time and resources.

TABLE 4. Sample size calculation example.

Standard Care Reinfarction Rate	Supervised Exercise Group Reinfarction Rate	Difference of Interest	*N* per group	Clinically Relevant Difference?
30%	25%	5%	1674	Probably not
30%	20%	10%	392	Maybe
30%	15%	15%	161	Likely
30%	10%	20%	82	Clearly

POWER IN NEGATIVE STUDIES

Accurate sample size calculations can help ensure the efficient use of scarce research resources by avoiding unnecessarily large studies or those so small that they have little probability of detecting important effects. Perhaps just as important, however, is that an appropriate sample size be based on well-justified, clinically important differences will enhance the strength of a negative study. That is, a negative study that has adequate power is more definitive than one in which sample size and power did not receive careful attention in the planning stage. A classic paper by Freiman and colleagues (10) illustrates the point that many apparently negative results are merely inconclusive due to inadequate sample sizes and low power.

STATISTICAL INFERENCE

The role of data analysis is to summarize the information collected in a research project. The summarization process begins with the computation of descriptive statistics and may involve graphical representation of the data. However, the most important summarization may be the conclusions that are reached based on the data. At the heart of the statistical inference process there are both logical issues and mathematical considerations.

WHAT IS A *P* VALUE?

The logical process involved in standard statistical inference begins with some reasonable assumptions about the nature of the data, how it has been collected, and an assumption that the "null hypothesis" is true. Generally, the null hypothesis states that there is no effect due to the intervention or factor of interest. For instance, if the outcomes are summarized by means, the null hypothesis is a mean difference of zero. Given these initial assumptions and the observed data, a statistical test is performed by computing a test statistic that reflects the effect of interest. The probability is computed for all possible test statistics that are at least as far away from the value posited by the null hypothesis as the one observed. This probability is the "*P* value" for the test. If the computed *P* value is below the pre-specified alpha level (e.g., 0.05), this is considered evidence against the assumption stated in the null hypothesis. In fact, the null hypothesis is rejected and the alternative or research hypothesis is accepted.

In a randomized trial of the effect of exercise training in patients with heart failure (8), one outcome was the change from baseline to 24 weeks in peak exercise oxygen consumption. In the control group, an increase of 58 ± 38 mL • min^{-1} was observed. For the exercise intervention group, the increase was 231 ± 54 mL • min^{-1}. The t-statistic is equal to the difference in means (231 − 58) = 173, divided by the standard error for the difference (66.9 in this example). The t-statistic therefore had a value of 2.59. Figure 1 illustrates how the *P* value for this hypothesis test, which is 0.0154, relates to the distribution of the t-statistic under the null hypothesis. That is, the area or probability under the t-distribution curve to the right of 2.59 and to the left of −2.59 is 0.0154. The quantity 0.0154 is the *P* value for this test, and since it is less than 0.05, the null hypothesis of no effect of the intervention on exercise duration is rejected in favor of the alternative of a beneficial effect.

CONFIDENCE INTERVALS

It is often more informative to report confidence intervals instead of just *P* values alone or *P* values with point estimates of the effect of interest. A confidence interval is computed so that one has a specified level of confidence that the quantity of interest is captured. For instance, a 95% confidence interval will include the quantity of interest, on average, 95 times out of 100. One useful interpretation of a confidence interval is that it is the range of values that are statistically consistent with the observed data. Confidence intervals usually have a close connection to the hypothesis test result, in that if the null hypothesis value is included in the confidence interval, the null hypothesis is not rejected. Conversely, if the null value is outside the interval, the test is significant and the null value is rejected.

To illustrate, the 95% confidence interval for the difference between the intervention and control groups for maximal exercise oxygen consumption in the paper by Keteyian et al. (8) was 37 to 309. As expected from the significant hypothesis test result, the null hypothesis value of zero is not included in this interval. In the same study, the difference in the change in SBP was only 1.0 mm Hg in favor of the intervention group, with a 95% confidence interval of −9 to 11. Therefore, the *t* test for SBP was nonsignificant with a *P* value of 0.842. Reporting both *P* values and point estimates along with confidence intervals is valid and complementary.

MULTIPLE COMPARISONS

The statistical inference process described above holds the probability of a Type I error to 5% or less for any *single* hypothesis test. However, there are many circumstances when a research project will need many hypothesis tests, and the tests may overlap in their implications. Statistical textbooks generally address multiple comparisons within the context of a one-way analysis of variance (ANOVA) when there are many means to be compared and many comparisons to be made among them. In this circumstance, there are several well-established options for doing the multiple comparisons, including the Scheffe test, Tukey's test, Fisher protected least significant difference (LSD), the Student-Newman-Keuls (SNK) and Duncan's multiple range tests, and Bonferroni (11). The Scheffe test tends to be very conservative, while Duncan's multiple range procedure may be too liberal. Fisher LSD and SNK methods are often useful but may be overly liberal when some cases allow the Type I error rate to exceed 0.05.

Although the multiple-comparison issue is most commonly associated with comparisons of means in a one-way ANOVA, it can arise in other circumstances. Consider a study of the effect of exercise on blood pressure in three groups: control (no exercise), aerobic exercise, and weight training. The study population may have multiple potential subcomponents. For instance, the effects for each sex, for

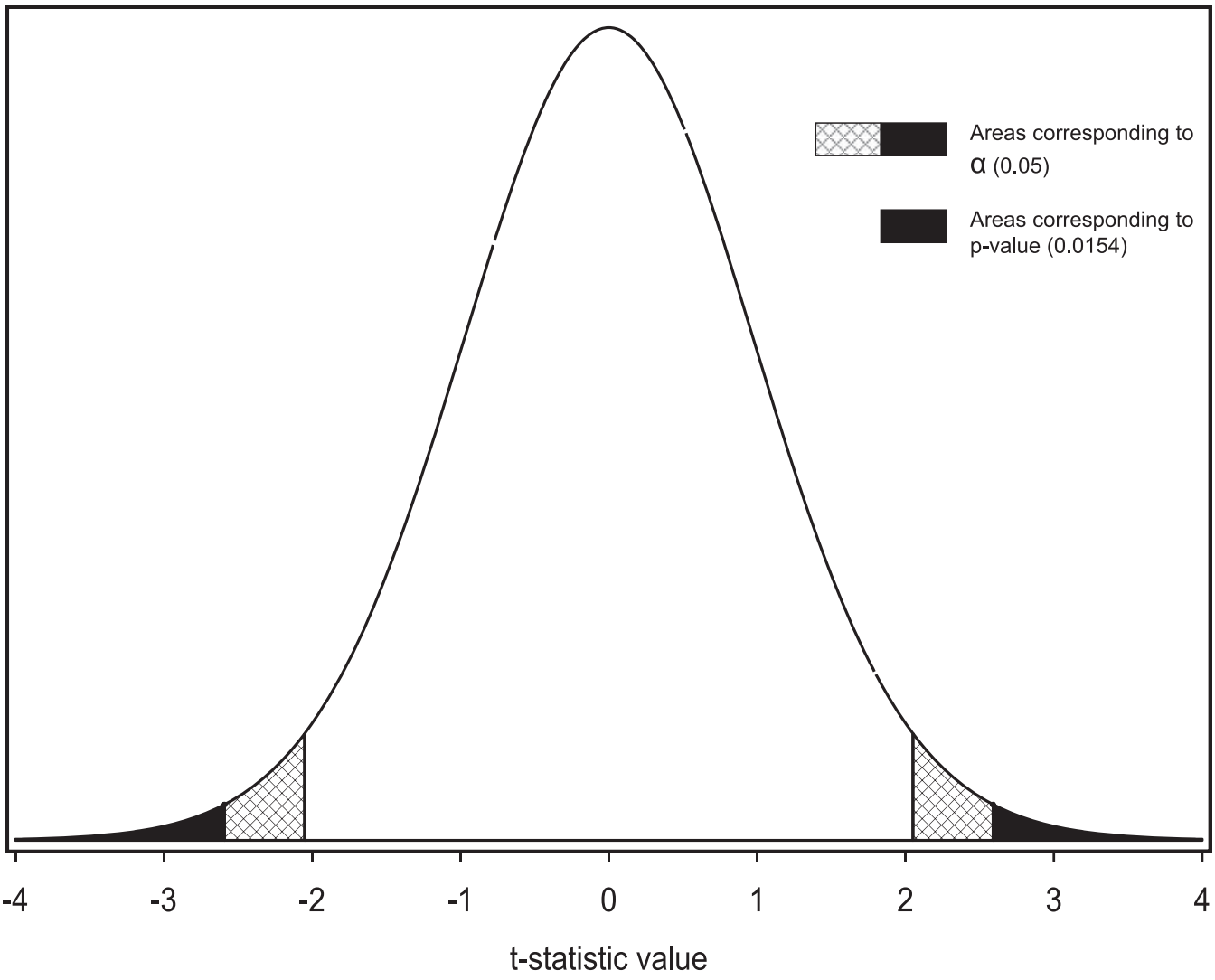

FIGURE 1. Example of *t* test *P* value for maximal exercise oxygen consumption.

black people and white people, and for different age groups (e.g., ages 50-64 and ages 65-75) may be of interest. If separate analyses are performed for each type of exercise and within every subpopulation, with the α level set to 0.05 for each, the probability of a Type I error among them will rise to something much higher than 0.05. In this example, with two types of exercise and eight subpopulations (i.e., white males ages 65-75, white males ages 50-64, etc.), the probability under null hypothesis that 1 *P* value out of the 16 will be under 0.05 will not be 1 in 20, but more than 1 in 2.

There are several approaches to maintaining a true overall α level in an experiment such as this. One technique is to use a much smaller critical level for each test so that the overall α level is held to 0.05. The simple Bonferroni adjustment is often used for this. In this example, each of the 16 α levels is set to be equal to 0.05/16 = 0.003125. There are variations on Bonferroni, such as Holm's approach (12), which can allow some of the tests to be performed with a critical α level greater than 0.003125.

Another major option is to approach the analysis in logical stages. For instance, an analysis might begin with a single test for an exercise effect using all the data, and if it is significant, only then are the many subgroups tested with α levels of 0.05. Alternatively, when a large number of related tests are possible, a smaller, prespecified subset can be identified for primary testing to use a less restrictive adjusted α level.

Finally, there are those who argue that no multiple comparison adjustments are required if all the tests are of interest in and of themselves (13-15). The strongest argument in favor of this position may be the apparent logical inconsistency that can result from considering the different conclusions that the same (sub) set of data would support, depending on how many other groups were included in a study design. However, regardless of how multiple-comparison questions are addressed, it is strongly recommended that multiple-comparison issues be discussed in the analysis and reporting of results, even if only to explain why no adjustment was undertaken.

SUMMARY

To make the best use of scarce research opportunities, researchers and those who make use of the research results

both need to be aware of basic statistical concepts as well as the techniques of statistical inference. An understanding of the strengths and limitations of hypothesis testing and confidence intervals is essential. Appropriate study design must start with a sample size calculation based on a meaningful effect of interest. With careful attention to statistical considerations, the value of a research project can be enhanced, both for researchers directly involved and for those who make use of the results. The second paper in this two-part series will describe basic analytical techniques and discuss some common mistakes in the interpretation of data or study results.

Acknowledgments: Originally published in Clinical Exercise Physiology 2001;3(3):121–126.

REFERENCES

1. Heath GW. Epidemiologic research: a primer for the clinical exercise physiologist. Clin Exerc Physiol. 2000;2:60–7.
2. Rosenbaum PR. Observational studies. New York: Springer-Verlag; 1995.
3. Schlesselman H. Case-control studies. Design, conduct, analysis. New York: Oxford University Press; 1982.
4. Hennekens CH, Buring JE. Epidemiology in medicine. Boston: Little, Brown & Co.; 1987.
5. Hill AB. The environment and disease: association or causation? Proc R Soc Med. 1965;58:295–300.
6. Piantadosi S. Clinical trials: a methodologic perspective. New York: Wiley & Sons; 1997.
7. Cohen J. Statistical power analysis for the behavioral sciences. Orlando, FL: Academic Press; 1977.
8. Keteyian SJ, Levine AB, Brawner CA, Kataoka T, Rogers FJ, Schairer JR, Stein PD, Levine TB, Goldstein S. Exercise training in patients with heart failure: a randomized, controlled trial. Ann Int Med. 1996;124:1051–7.
9. Elashoff JD. nQuery Advisor version 2.0 user's guide. Los Angeles: Dixon Associates; 1997.
10. Freiman JA, Chalmers TC, Smith I Jr, Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. Survey of 71 "negative" trials. N Engl J Med. 1978; 299:690–4.
11. Miller RG. Simultaneous statistical inference. New York: Springer-Verlag;1981.
12. Holm S. A simple sequentially rejective multiple test procedure. Scand J Statistics. 1979;6:65–70.
13. Poole C. Beyond the confidence interval. Am J Public Health. 1987;77:195–9.
14. Poole C. Confidence intervals exclude nothing. Am J Public Health. 1987;77:492–3.
15. Rothman K. Modern epidemiology. Boston: Little, Brown & Co.; 1986.