REVIEW

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

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

ABSTRACT

This paper is the second in a two-part series intended to provide a brief overview of some of the important concepts in the field of biostatistics. In this paper, basic analysis methods are reviewed and issues important to the conduct and interpretation of research studies are discussed. Statistical methods of analysis are dependent on the type of data to be analyzed. Five common types of data are briefly explained: nominal, binary, ordinal, discrete, and continuous. Basic analysis methods are presented in the context of these defined data types. The interpretation of a study's results is the final critical step upon completion of a research study. The issues of critical thinking, bias, confounding, validity and the potential for over-interpretation of research results are discussed. Understanding biostatistical concepts and appropriately employing them over the course of the study is an essential part of quality research. *Journal of Clinical Exercise Physiology*. 2018;7(4):94–103.

Keywords: study design, statistical analysis, validity

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-07-02 via free access

INTRODUCTION

In the first part of this two-part series, issues important in the design of exercise physiology research were discussed, along with the basic concepts underlying statistical inference (1). In this second part, basic statistical methods and interpretation of analysis results are reviewed, including further discussion of the importance of power and its implications in the reporting of negative studies. Additionally, the importance of critically evaluating research results for potential biases are described. Table 1 provides brief definitions for some of the statistical terms used in this paper (2,3,4).

ANALYSIS TECHNIQUES

The analysis techniques should fit the study design and the types of data collected. After identifying the type of data for both the outcome (dependent) variable and the predictor (independent) variable, the proper analysis method can be selected. Five common data types are nominal, binary, ordinal, discrete, and continuous (Table 2).

Types of Data

Nominal or categorical types of data are data that are grouped but have no inherent ordering to them. For instance, the variable identifying treatment group (e.g. exercise vs. non-exercise) is considered a nominal variable. Hand dominance, as in right-handed, left-handed, and ambidextrous, would be another example of nominal data. *Binary* data is a special case of nominal data in that it consists of only 2 categories. For example, biological sex is a binary variable.

Ordinal variables are slightly more complicated than nominal in that they are categorical variables *with* an inherent ordering. The New York Heart Association functional classification of congestive heart failure is an example of an ordinal type of variable. Class I describes a patient who is not limited in normal physical activity by symptoms. Class II occurs when ordinary physical activity results in fatigue, dyspnea, or other symptoms. Class III is characterized by marked limitation in normal physical activity, while Class IV is defined by having symptoms at rest. Therefore, a Class II designation signifies more advanced heart failure than Class I, but it is unknown if the difference between Class I and Class II is the same as the difference between Class II and Class III.

This potential dissimilarity of the magnitudes of differences between the outcome values is what differentiates ordinal data from *discrete* or *continuous* data. For these latter

[1]Department of Public Health Sciences, Henry Ford Hospital, One Ford Place, Suite 3E, Detroit, MI 48202 USA

Address for correspondence: George W. Divine, PhD, Department of Public Health Sciences, Henry Ford Hospital, One Ford Place, Suite 3E, Detroit, MI 48202; (313) 874-6724; Fax: (313) 874-6730; e-mail: gdivine1@hfhs.org.

Conflicts of Interest and Source of Funding: None.

Copyright © 2018 Clinical Exercise Physiology Association

TABLE 1. Statistical terms.

Term	Definition
Variable	A name for measurements in a study
Dependent or Outcome Variable	A variable that is identified as an effect, result, or outcome. The dependent variable is sometimes conceptually viewed as being caused by the independent variable
Independent or Predictor Variable	A variable that is identified as a possible causal variable
Covariate	Additional predictor variable. Often of secondary interest
Normal Distribution	Theoretical distribution of values that is symmetrical, unimodal, and bell-shaped
Parametric Methods	Used when the data studied are from a sample or population that is normally distributed. Data should be discrete or continuous
Nonparametric Methods	Used when the data studied are from a sample or population that is not normally distributed
Univariable Analysis	Analysis involving only 1 variable, e.g. paired *t* test
Bivariable Analysis	Analysis involving 1 dependent variable and 1 independent variable. Also commonly referred to as univariate analysis
Multivariable Analysis	Analysis involving 1 dependent variable but 2 or more independent variables
Multivariate Analysis	Analysis involving more than 1 dependent variable
Relative Risk/Rate Ratio	Risk of outcome when the factor of interest is present Risk of outcome when the factor of interest is not present Note: Typical measure in cohort studies
Odds Ratio	Odds of outcome when the factor of interest is present Odds of outcome when the factor of interest is not present Note: Typical measure in retrospective and cross-sectional studies. An odds ratio can provide a reasonable estimate of the relative risk in certain circumstances
Blinding	Being unaware of certain aspect(s) of the study. Typically, a single-blind study is when subjects in the study are unaware of which treatment they receive. A double-blind study is where both subject and study staff are unaware of which treatment is being administered
Validity	Lack of systematic error
Internal Validity	The validity of the inferences drawn as they pertain to the source population
External Validity (Generalizability)	The validity of the inferences drawn as they pertain to people outside the source population. Internal validity is a prerequisite for external validity
Bias	Any process at any stage of inference that tends to produce results or conclusions that differ systematically from the truth
Biologic Plausibility	A known biological mechanism that supports the hypothesis of interest
Confounding	When a third, possibly unsuspected, variable changes the apparent association between the study outcome and the factor of interest because of its relationship to both

TABLE 2. Types of data

Type of Data	Definition	Common Example(s)
Nominal (Categorical)	Unordered categories	Race: Black/White/Other
Binary	Two categories; Special case of nominal	Yes/No; True/False; Male/Female
Ordinal	Ordered categories	Functional symptom classification: None/Mild/Moderate/Severe
Discrete	Ordering and magnitude important; restricted to specific values (usually integers)	Number of times subject has a MI; days per week exercised
Continuous	Measurable quantity; not restricted to specific values	Age; serum cholesterol; Peak VO_2

MI = myocardial infarction; VO_2 = volume of oxygen consumed

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-07-02 via free access

TABLE 3. Common statistical hypothesis testing methods and some of their characteristics.

		Type of Data Required				
Test	**Number of Groups**	**Outcome Variable**	**Independent Variable**	**Compares, Tests or Estimates**	**Null Hypothesis**	**Assumptions**[a]
Paired *t* test	1	Continuous	Categorical	Mean Difference	Mean Difference = 0	Normality; Paired data
Wilcoxon Signed Rank Test	1	Continuous Discrete or Ordinal	Categorical	Center of Difference Distribution	Median Difference = 0	Symmetric distribution of differences
McNemar Test	1	Categorical	Categorical	Discordant Pairs	A discordant pair is as likely to go in one direction as the other	Paired data
Pearson's Correlation	1	Continuous	Continuous	Linear Association	The correlation is 0	Normality, Linear Relationship
Spearman's Correlation	1	Continuous Discrete or Ordinal	Continuous Discrete or Ordinal	Linear Association	The correlation is 0	Independent and identically distributed
Simple Linear Regression	1	Continuous	Continuous	Slope and Intercept	Slope = 0	Normality, Linear Relationship
Two-Sample Student *t* Test	2	Continuous	Categorical	Means	Mean Difference = 0	Normality; Equal variances
Wilcoxon Rank Sum Test	2	Continuous Discrete or Ordinal	Categorical	Distributions	Prob(x < y) = 0.5 (The probability that an observation from group 1 is less than an observation from group 2 is 50%)	Distributions have same shape
χ^2 / Fisher Exact Test	≥ 2	Categorical	Categorical	Proportions	The proportions in all categories are equal	Independent and identically distributed
χ^2 Test for Trend	≥ 2	Discrete or Ordinal	Categorical	Trend in Proportions	The proportions in all categories are equal	Independent and identically distributed
Analysis of Variance (ANOVA)	≥ 3	Continuous	Categorical	Means	All the means are equal	Normality; Equal variances
Kruskal-Wallis Test	≥ 3	Continuous Discrete or Ordinal	Categorical	Distributions	Prob(x < y) = 0.5 (The probability that an observation from group i is less than an observation from group j is 50%)	Distributions have same shape

[a] All the tests shown require an assumption of independent observations and random sampling

2 data types, a unit change has the same interpretation no matter where it is on a scale. Examples of discrete data, most often count data, are length of hospital stay (in days only) or *counts* of events such as number of emergency department visits in the past year. Continuous type data examples are weight, height, age, and many lab measurements.

Described below are common bivariable analysis methods for each data type, listed by outcome variable type. By definition, bivariate analysis is used when there is 1 outcome variable and 1 independent variable. Table 3 briefly describes some of the characteristics of the most common testing methods. More detailed guidance on how to select the most appropriate analysis technique is available (2).

CONTINUOUS OUTCOME VARIABLE

Two Sample Student t Tests

The 2-sample Student *t* test is probably the most commonly used statistical technique. This test is used with continuous,

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-07-02 via free access

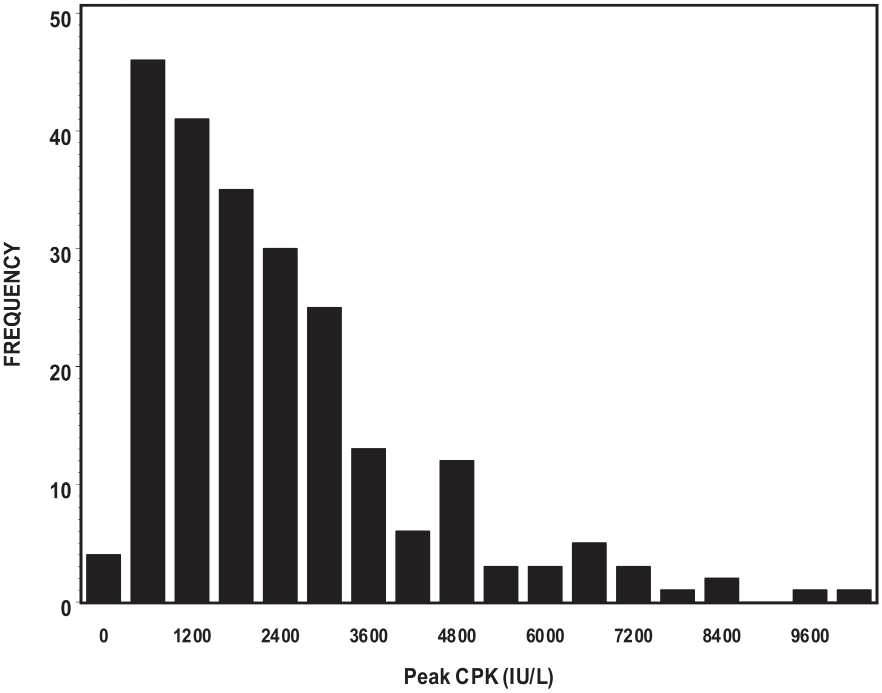

FIGURE 1. Creatine Phosphokinase (CPK) Levels in 231 male patients hospitalized for first myocardial infarction (MI). The data appear to be from a skewed, non-normal distribution. These data would typically require transformation or analysis by non-parametric techniques.

normally distributed variables from 2 independent groups with equal variances. It is selected when relating a continuous outcome variable with a binary independent variable. For example, if the study consists of an exercise regimen group and non-exercise group, and compares the change in HDL cholesterol between the 2 groups, a 2-sample *t* test is often the appropriate test. The hypothesis tested by the *t* test is whether the mean change in HDL cholesterol is the same in the 2 groups.

One circumstance where the 2-sample *t* test would not be appropriate is if the data were dependent (not independent). For example, in a single sample prospective cohort study, HDL cholesterol measurements are taken on a group of sedentary individuals at baseline and again after a 4-month exercise regime. The correct test to compare the average change in HDL would be a paired *t* test, as the data were measured before and after exercise training within the same group of people. A strength of the paired design is that each individual serves as their own control, which can reduce variance and, in turn, increase the chance of getting a statistically significant result, if a difference really exists.

Another circumstance where a Student *t* test is not appropriate is if the data are not normally distributed. One example would be the distribution of peak creatine phosphokinase (CPK) values for myocardial infarction (MI) patients as shown in Figure 1. For the analysis of such data, transformations or nonparametric testing methods should be employed. Common examples of data transformations are the square root, the log, and the inverse. Minor inequality of variances can be corrected using a modified form of the *t* test, such as the Welch test.

Analysis of Variance

One-way analysis of variance, or ANOVA, is the extension of the Student *t* test when more than 2 groups are compared. Again, the assumptions of continuous, normally distributed data from independent groups of data with equal variances are important here. ANOVA is selected when relating a continuous outcome variable with a nominal independent variable. An example of when it is appropriate to use ANOVA is in a trial where individuals are randomly assigned to 3 groups such as usual care (controls), usual care + mild exercise, and usual care + moderate exercise. An outcome measurement might be change in resting heart rate, which is a continuous and often normally distributed variable. The hypothesis tested by the one-way ANOVA is that the average change in resting heart rate is the same across the 3 groups.

Correlation

The two analysis methods discussed so far deal with tests of association between a continuous outcome variable and a nominal independent variable. When both variables (x and y) are continuous, the association is assessed by computing their correlation coefficient or by linear regression analysis. A correlation coefficient, denoted as *r*, expresses the extent to which when x is large, y is large, or when x is small, y is small. The correlation coefficient can take on values between −1.0 and 1.0, with 1.0 indicating perfect positive correlation and −1.0 indicating perfect negative correlation. The usual null hypothesis is that the correlation is 0, which indicates no (linear) association between x and y. That is, under the null hypothesis, knowing either x or y from a pair tells one

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-07-02 via free access

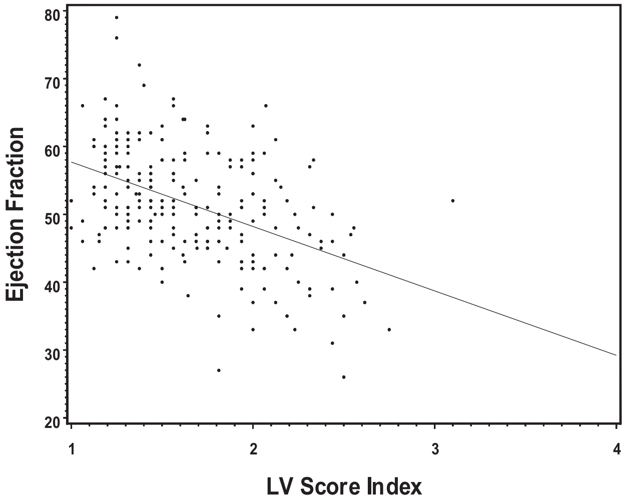

FIGURE 2. Example of correlation between Left Ventricular (LV) Score Index and Ejection Fraction in 231 male patients hospitalized for first myocardial infarction (MI) ($r = -0.46$).

nothing about what the corresponding y or x value is likely to be. Figure 2 shows the relationship between ejection fraction and left ventricular (LV) score index in 231 male patients hospitalized for a first MI. In this example, the correlation is −0.46. This means there is a moderate relationship between low LV score index and high ejection fraction.

Regression

The association between x and y can also be expressed in a linear model that takes the form of a straight line: $y = b_0 + b_1x$, where b_0 represents the intercept and b_1 the slope. There is a very close relationship between a correlation analysis and linear regression. The null hypothesis that the slope is 0 is mathematically equivalent to the hypothesis that the correlation is 0, and the significance tests and *P* values are the same. Figure 2 shows the regression line for ejection fraction as a function of LV score index. The regression equation is Ejection Fraction = 67.2 − 9.5 (LV score index). That is, the average expected ejection fraction decreases by 9.5% for a 1.0 unit increase in LV score index.

NOMINAL OUTCOME VARIABLE

χ^2 Test

A χ^2 test is a commonly used test when data are grouped into categories. In particular, both the outcome and independent variables are nominal. For example, the most common type of categorical data are items measured with a *yes* or *no* response, such as when testing for the frequency or occurrence of an event during a study. In comparing outcomes in an exercise group with outcomes in a non-exercise group, χ^2 tests would be used to compare the occurrence of shortness of breath. The data may also consist of more than two categories. For instance, if subjects are asked what time of day their shortness of breath was most bothersome and the possible responses are morning, afternoon, or night, a χ^2 test can then be used to determine if there are differences in the typical time of day distribution between study groups. The χ^2 test defined here is known as the χ^2 test of independence or χ^2 test of association.

The Fisher's exact test is used in place of the χ^2 test when an outcome and/or exposure is rare. The mathematical rule is that if the expected value of a cell is less than 5, a Fisher's exact test should be used. Commonly, this can happen when 1 or more of the cell sizes are less than 10. A McNemar test is used in the paired data situation when there are 2 nominal variable types.

χ^2 Test of Trend

A nominal outcome variable with an ordinal independent variable can be tested with a χ^2 test of trend. For example, if the independent variable is discomfort upon exercise with categories mild, moderate, and severe, and the outcome is adherence to exercise regimen (yes/no), then a reasonable hypothesis to look at might be, does adherence to exercise regimen decrease with increasing levels of discomfort? The most appropriate statistical test for this hypothesis is the χ^2 test of trend.

ORDINAL OUTCOME DATA

Nonparametric Testing

Nonparametric tests are used when the data do not fit the assumption of continuous, normally distributed variables with equal variances in separate groups. In particular, nonparametric tests are selected when testing ordinal outcome variables (Table 3). The Wilcoxon rank sum test is the nonparametric alternative to the Student *t* test. The alternative to the paired *t* test is the signed rank test, and the alternative to ANOVA is the Kruskal-Wallis test. There is also a nonparametric alternative to Pearson's correlation called Spearman's correlation. Each of these nonparametric methods is based upon ranking the data, and computing test statistics based upon those ranks. The lower sensitivity of ranks to outliers and other distributional characteristics is a major reason nonparametric methods are suitable when techniques with parametric assumptions do not hold. Applying parametric analysis techniques to data that does not fit the parametric assumptions can be improper and misleading.

Discrete Data

Analyzing discrete data can pose unique challenges. If discrete data are normally distributed, analysis methods that depend on parametric assumptions can be selected. Often, however, the normality assumption is not met, and as a result, nonparametric tests must be used. A third option when dealing with discrete data is to categorize the data before analyzing it. For example, the number of prior MIs is a discrete outcome variable that is often of interest. Due to scarcity of data at the high end, it may be necessary to categorize as *none*, *1*, and *2 or more*.

MULTIVARIABLE ANALYSES

The above analysis sections have only described bivariable analyses, that is, when there is 1 outcome variable and 1 independent variable. When there is more than 1

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-07-02 via free access

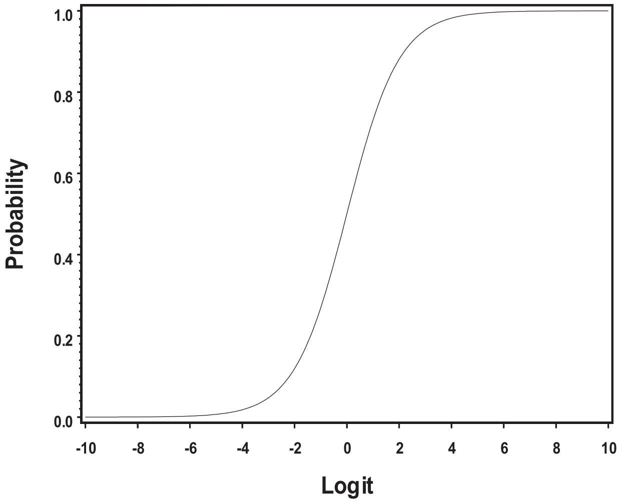

FIGURE 3. Illustration of the relationship between a probability (p) and its logit. $\text{logit}(p)=\log(p/(1-p))$.

independent variable to consider, multivariable analyses such as multiple linear regression, multiple logistic regression, or Cox regression need to be selected. When there is more than 1 *outcome* variable to be considered, multivariate methods such as multivariate ANOVA (MANOVA) can be selected. It is beyond the scope of this presentation to describe multivariate methods. Multivariable analysis can have many uses, including modeling, prediction, and testing for associations while taking into account other variables. There are numerous considerations to be taken into account for each of these applications. The presentation here will be limited to a brief description of a multiple logistic regression example.

Logistic Regression

Standard linear regression cannot be used to analyze binary outcomes. Since the outcome variable can only take on 2 possible values, it cannot meet the normal distribution assumption. In addition, extreme values of the independent variable(s) will result in *impossible* predicted probabilities that will be greater than 1.0 or less than 0.0.

To address these issues, instead of modeling the probability of an outcome directly, a *logistic* model is fit. That is, a regression model predicting the *logit* of the probability of interest is fit to the data. In this case, a multiple logistic regression model would have the form:

$$\text{logit}(p)=\log\left(\frac{p}{1-p}\right)=\beta_0+\beta_1 x_1+\cdots+\beta_n x_n$$

where p is the probability of the event being modeled. As can be seen from Figure 3, for extreme values of the independent variable(s) the predicted logit values will asymptotically approach 0.0 or 1.0. Noting that $p/(1-p)$ is the odds of the outcome of interest, it can be seen that each of the coefficients in the logistic equation gives the log of the change in the odds ratio for a unit change in the associated predictor variable.

For an example of a multiple logistic regression model, consider an analysis to assess the association between race and a prior diagnosis of hypertension among 328 men diagnosed with a first MI. Among 241 white patients, 98 (41%) reported a prior diagnosis of hypertension, compared to 47 of 87 (54%) among black patients. This would give an estimated odds ratio of 1.72 for hypertension for black MI patients ($P=0.032$). However, age is also significantly associated with both hypertension and race. The average age in years for hypertensives was 59.4 versus 56.0 for non-hypertensives. For black patients the average age was 60.1 versus 54.3 for whites. Since age is associated with both race and hypertension, it is a potential confounder.

A multiple logistic regression model can be used to estimate the odds ratio for race and hypertension, adjusting for age. For such a model, the coefficients are 0.449 and 0.034 for black race and age, respectively. Expressed as odds ratios (by taking the anti-log for each coefficient), these correspond to estimates of 1.57 and 1.03, respectively. The odds ratio after adjustment for age is still greater than the null hypothesis value of 1.0, but it is reduced and the P value is no longer quite significant ($P=0.083$).

It is important to note that the odds ratio for age and hypertension of 1.03 is for each additional year of age. It is often convenient to rescale a variable like age in decades, so that the estimated odds ratio will have more significant digits. In this case, the odds ratio for each additional decade of age is 1.40.

INTERPRETATION OF RESULTS

Interpreting results is just as important as collecting high-quality data and doing the correct analysis. Described next are important issues to take into account after data has been collected and the basic analyses have been performed and summarized.

Failure to Use Critical Thinking

Although it may not qualify as a single mistake, the most common error in data interpretation is probably a failure to think critically. Colton (5) illustrates 9 types of data misinterpretation or "Fallacies in Numerical Reasoning." In 1 of Colton's examples, the reader is asked to consider an assertion that the common association of heart attacks with strenuous activity may be misleading, because only 2% of heart attacks occur during exercise and over 50% occur during sleep or while otherwise at rest. Colton notes that most heart attack victims spend much less than 2% of their time exercising, so the heart attack rate per unit of exercise time will still be much higher than for sleep or rest if the appropriate denominator is taken into account. If one were to survey MI patients and a control population about their physical activity, it would be important to be certain the responses applied to levels before the infarct. Failing to take into account the proper temporal association could result in erroneous conclusions about the strength of the association between activity and the risk of heart disease.

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-07-02 via free access

Colton (5) and Roht and colleagues (6) both give many examples of basic misinterpretations of data that might be avoided by critical thinking.

Bias

Although most misinterpretations can be recognized and avoided by careful thought, a more systematic approach to considering such potential problems can be helpful. Heath (7) describes 8 types of bias that can affect epidemiologic studies when the data is collected. More general forms of bias can result from broader study design and interpretation issues. For instance, Sackett (4) lists 57 different biases in 6 categories.

In its most basic form, bias may be due to the way in which observations are made. For instance, the routine clinical measurement of blood pressure is subject to a large *digit preference* bias, where a reading such as 120/80 mmHg is much more common than one of 122/78 or 118/82. If this bias is just as likely to go in one direction as another, such measurements might satisfy the general definition of being accurate. That is, on average, they could give the correct blood pressure. However, if blood pressure is only measured to the nearest 10 mmHg, this would lack precision compared to the more appropriate nearest even integer. That is, such measurements would have only one-fifth of the potential precision. Figure 4 illustrates the difference between precision and accuracy. Ideally, measurements can be made with both precision and accuracy. For instance, study personnel can be trained to measure blood pressure accurately and without digit preference, and their performance can be verified by testing.

Finally, selection bias is very common in uncontrolled studies. For instance, when a new treatment is first tested, it might be ethically prudent to try and use relatively healthy patients who might best withstand unknown deleterious effects. More simply, if the usual medical practice of *selection by indication* is used, a case series will over-represent subjects who are most likely to benefit or who are most likely to do well.

When biases cannot be eliminated through their control by the study design, such as using a randomized control group and blinding, they can often be reduced by taking them into account when the data is analyzed. The section below on multivariable adjustment illustrates one way a bias due to confounding can be addressed.

Confounding

In a critical thinking example, Roht and colleagues (6) ask the reader to evaluate the observation that “judging by death rates of 1950 in the U.S., the safest occupation for men is messenger boy.” Although one might attribute such an observation to the protective effect of a messenger's high level of physical activity, a more basic explanation would be the young age for most messenger boys and the low death rates associated with youth. When both the study outcome and the factor of interest are associated with an important third factor, *confounding* will be present. That is, an association between the factor of interest and the outcome can be introduced or masked because of the association with the confounder. In this example, and in the race and hypertension example given earlier, age was a confounder. Confounding can be avoided by a study design in which the groups to be compared are balanced with respect to potential confounders. For observational studies in which balance cannot be controlled as part of the design, it is often possible to make a statistical adjustment to correct for an imbalance. The latter was the approach illustrated using logistic regression for the race and hypertension example.

Internal Validity

Maintaining comparability among all subjects and measurements in a study will preserve *internal validity*. For example, the Veterans Administration Cooperative Study on Antihypertensive Agents (8) used an extensive placebo run-in period which was monitored to ensure the study population consisted of individuals who would actually take the prescribed study medication. Having clear and uniform standards for making observations and collecting data is another critical factor for preserving internal study validity. Finally, the potential for a lack of internal validity is one of the major reasons that results from retrospective or *uncontrolled* cohort studies should be interpreted carefully.

External Validity (Generalizability)

Comparability of a study population to the population at large is usually required for *external validity*, where study results may be extrapolated to people or patients in general. There can be a trade-off between internal and external validity. For instance, in order to get as good an assessment of a treatment effect as possible, it is often the case the clinical trials are restricted to patients with greater severity or a particular time in the course of their disease. For instance, the National Institute of Neurological Disorders and Stroke trial of t-PA for stroke restricted enrollment to patients within 3 hours of stroke onset (9). Although this allowed a definitive conclusion about the benefit of t-PA for stroke patients treated this early, it is not generalizable to the larger population of stroke patients who present later.

Biologic Plausibility

Consideration of biologic plausibility can be an important issue to be addressed in the discussion section of a research paper. The analysis should generally be presented objectively without undue influence from outside expectations. When the results and their implications are discussed more broadly, assessment of how well they fit with other knowledge and theory is appropriate, if not essential to the success of the study.

Over-Interpreting Positive and Negative Study Results

When a study is complete and appropriate analysis methods have been applied, there are still issues to be addressed in reporting the results. When formal statistical hypothesis

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-07-02 via free access

REVIEW

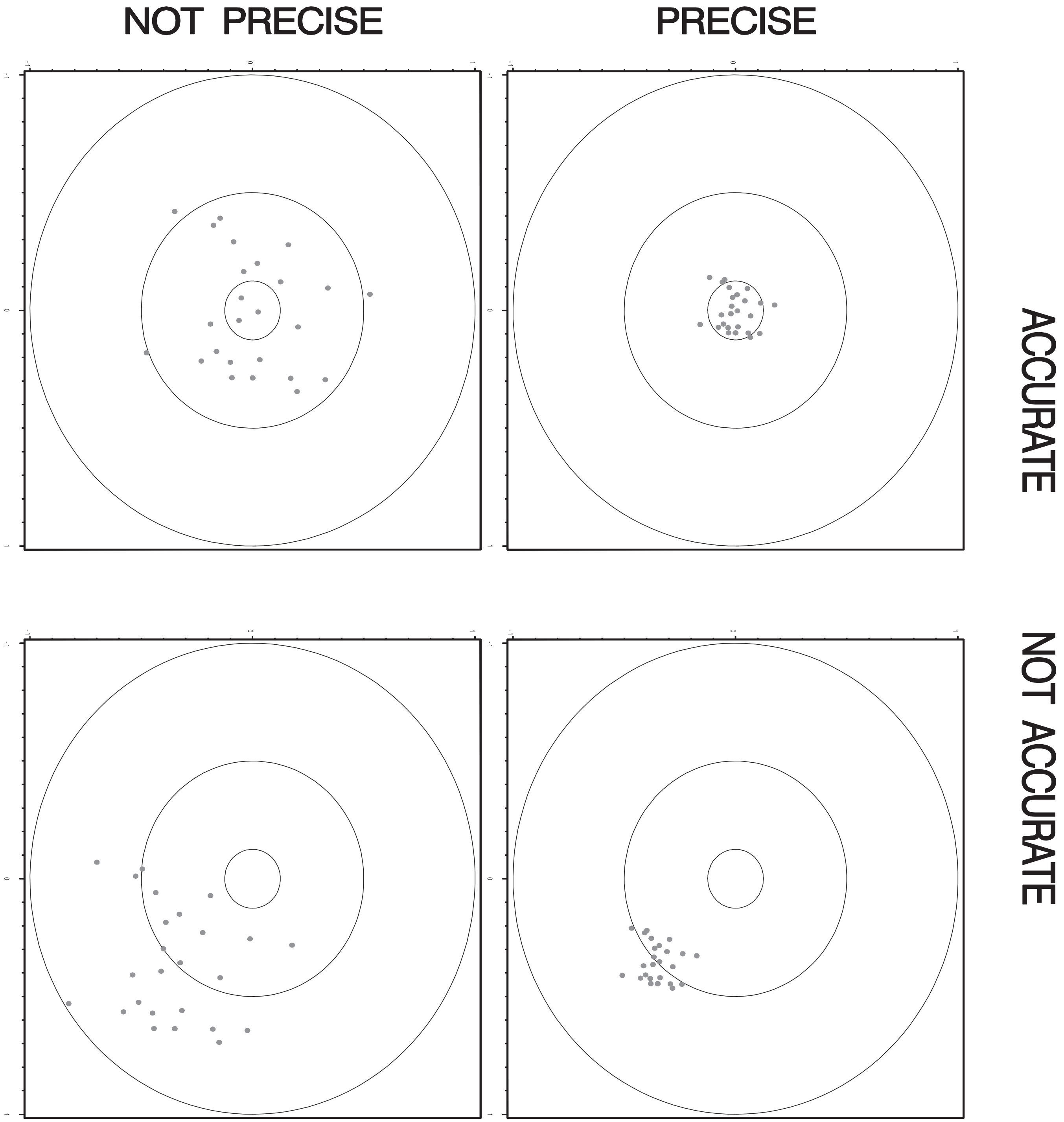

FIGURE 4. Precision alone can be thought of as a tight cluster of observations, with little spread among them, with the center not on the center mark, which represents truth. Accuracy alone occurs when the average of the points, no matter how spread out, end up in the center, representing truth. A desirable measure, hence, has both precision and accuracy.

testing has been conducted, there are three main possible conclusions that can be reached: (a) The null hypothesis (Ho) has been rejected and therefore the factor of interest is associated with the outcome. (b) The Ho is not rejected, and there is evidence against a useful association. (c) The Ho is not rejected, but there is not enough evidence to rule out an important association.

When a study gives negative results, the presentation and interpretation should be approached carefully. A negative result may mean that the study has essentially *proven* the null hypothesis, or it may merely mean that not enough information was obtained to reach a definitive conclusion. Choosing between these two alternatives is important, especially when no formal sample size or power calculation was performed in the planning phase of a project. This is common when data from an existing dataset is analyzed to address a new question or a question that is secondary to the original goals of the project.

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-07-02 via free access

In this situation it can be extremely useful to compute a confidence interval for the basic summary measure that was not statistically significant. Since a confidence interval contains all the values that are statistically consistent with the observed data, if the confidence interval is narrow and excludes whatever the minimal clinically important effect should be, the study can be regarded as giving a definitive negative result. For example, according to a report from the Physician's Health Study, the relative risk of a cardiovascular event associated with having periodontal disease is 1.01 with a confidence interval of 0.88 to 1.15, after adjustment for important covariates such as level of physical activity (10). Since an excess risk of a 15% or less may not be of much practical importance, this should probably be regarded as a definitive negative result, subject only to the potential limitations for a non-randomized cohort study.

The second alternative for a negative result, where the confidence interval is wide, might be illustrated by the results of the MRFIT trial, when the main study report gave an estimated 7.1% reduction in mortality from coronary heart disease (CHD), with a 90% confidence interval of −15% to 25% (11). Since a 25% reduction in CHD mortality might be clinically useful, this negative result, while very discouraging, was not quite definitive, since it was still consistent with a clinically useful treatment effect.

Another way to assess the strength of a negative result could be to do a post hoc sample size or power calculation. Unfortunately, there are several ways that such calculations are performed, and the easiest ones may be uninformative or misleading. The least useful calculation is to present the *power* for the data and analysis just performed. Some data analysis software packages such as SPSS (12) automatically calculate power when statistical tests are performed. Unfortunately, the power that is automatically computed is for the particular effect that was contained in the data analyzed. This effect need not coincide with the minimal clinically important difference. Another limitation of such a power estimate is that it is basically a just one-to-one function of the *P* value for the test performed.

It would be more informative to report the sample size that would give 80% or 90% power for the observed effect. However, it would be most useful to employ the observed variability and compute the power to detect the minimum clinically important difference using the sample size of that study. In the case of the MRFIT trial, even though a careful sample size calculation was done at the planning stage of the study, a recalculation of the power taking into account the lower control group death rate, yielded a power of only 60% (13).

Confusing Statistical Significance and Clinical Significance

The preceding section discussed the situation where a researcher might be misled by a data analysis result of *no (statistical) significance* and conclude the results imply no clinical or practical significance.

Conversely, an extremely large sample size can result in very low *P* values for observed effects that may have little or no importance. For example, if a large observational study comparing active versus inactive individuals had a sample size of 3,100 per group, a blood pressure difference of 2.0 mmHg would be statistically significant ($P<0.0001$). However, the 95% confidence interval would range from 1.0 to 3.0 mmHg. Thus, although such a blood pressure difference would be statistically significant, at the same time it would rule out most differences of any real practical importance.

A more concrete example might be the CAPRIE trial results, where compared to aspirin, clopidogrel reduced the relative risk of stroke, MI, or vascular death by 8.7% ($P=0.043$) with a 95% confidence interval of 0.3% to 16.5% (14). Although the *P* value was significant and the results were reported as positive, others have commented that the benefit of clopidogrel is modest, considering its expense in comparison to aspirin (15).

SUMMARY

After an appropriate design and sample size have been selected, when implementing a research study it is important to proceed with care in collecting the data and in selecting the appropriate data analysis. Critical thinking is also required when interpreting and presenting the results of the study. Negative study results in particular should be interpreted carefully. With diligent attention to the design, implementation, analysis and interpretation of research studies, researchers can make the best use of scarce research resources.

Acknowledgments: Originally published in Clinical Exercise Physiology 2001;3(4): 4-12.

REFERENCES

1. Havstad SL, Divine GW. Biostatistical analysis: A primer for clinical exercise physiology, part I. Clin Exerc Physiol. 2001;3(3):121-6.
2. Riegelman RK, Hirsch RP. Studying a study and testing a test. Boston: Little, Brown and Company; 1996.
3. Rothman KJ, Greenland S. Modern epidemiology. Philadelphia: Lippincott-Raven Publishers; 1998.
4. Sackett, DL. Bias in analytic research. J Chron Dis 1979; 32:51-3.
5. Colton T. Statistics in medicine. Boston: Little, Brown and Company; 1974.
6. Roht LH, Selwyn BJ, Holguin Jr AH, Christensen, BL. Principles of epidemiology: a self taught guide. New York and London: Academic Press; 1982.
7. Heath GW. Epidemiologic research: a primer for the clinical exercise physiologist. Clin Exerc Physiol. 2000;2(2):60-7.
8. Veteran Administration Cooperative Study Group on Antihypertensive Agents. Effects of treatment on morbidity in hypertension. JAMA. 1967;202(11):116-22.
9. The National Institute of Neurological Disorders and Stroke t-PA Stroke Study Group. Tissue plasminogen activator for acute ischemic stroke. N Engl J Med. 1995;333(24):1581-87.

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-07-02 via free access

10. Howell TH, Ridker PM, Ajani UA, Hennekens CH, Christen WG. Periodontal disease and risk of subsequent cardiovascular disease in U.S. male physicians. J Am Coll Cardiol. 2001;37(2):445-50.
11. Multiple Risk Factor Intervention Trial Research Group. Multiple risk factor intervention trial. JAMA. 1982;248 (12):1465-77.
12. IBM SPSS Statistics Base 24. Version 24.0.0 [software]. IBM. 2018 August 30. [cited 2018 Jul 10]. Available from: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/24.0/en/client/Manuals/IBM_SPSS_Statistics_Base.pdf
13. Gotto AM. The multiple risk factor intervention trial (MRFIT): a return to a landmark trial. JAMA. 1997;277(7):595-7.
14. CAPRIE Steering Committee. A randomized, blinded, trial of clopidogrel versus aspirin in patients at risk of ischaemic events (CAPRIE). Lancet. 1996;348(9038):1329-39.
15. Baigent C. Clopidogrel reduced stroke, MI, and vascular death compared with aspirin. ACP Journal Club. 1997;126:59.

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-07-02 via free access