# Clinicians' Guide to Statistics for Medical Practice and Research: Part II

**Marie A. Krousel-Wood, MD,*† MSPH, Richard B. Chambers, MSPH,* Paul Muntner, PhD†**

**Ochsner Clinic Foundation, New Orleans, LA*
*† Department of Epidemiology, Tulane University School of Public Health and Tropical Medicine, New Orleans, LA*

In the last issue of the *The Ochsner Journal*, Part I of this series provided basic statistical principles for the practicing clinician to use in the review of the medical literature and in conducting clinical research. We included discussion of formulating the study question and study goals, choosing the appropriate study design, and using the appropriate statistical test (1). In this issue, we present the final part of the two-part series and will describe key issues relevant to statistical significance, statistical power, error, bias and confounding, and clinical relevance of study results. As with Part I, this is not intended to be a comprehensive review of statistical principles, but rather a quick reference for practical application of statistical principles.

*Address correspondence to:*
*Marie A. Krousel-Wood, MD, MSPH*
*Director, Center for Health Research*
*Ochsner Clinic Foundation*
*1514 Jefferson Highway*
*New Orleans, LA 70121*
*Tel: (504) 842-3680*
*Fax: (504) 842-3648*
*Email: mawood@ochsner.org*

*Richard Chambers performed this work at Ochsner prior to his employment at Pfizer Global Research and Development, 235 E. 42nd Street, New York, NY 10017.*

## Are the Results of the Study Significant? The Role of Chance

Given that the study question is relevant, and the appropriate study design and statistical tests were employed, the significance of the study results is usually of keen importance to the practicing physician and the clinical investigator. One typically directs his/her attention to statistical significance (often denoted by the "p-value" described below). However, statistical significance in cardiovascular and other studies does not always imply clinical relevance (2). The user of the information should be aware of this prior to incorporating the study results into his/her decision-making processes. When assessing statistical significance, it is important to evaluate the role of chance through hypothesis testing (or the performance of tests of statistical significance) and the estimation of the confidence interval. In discussing hypothesis testing, an explanation of error is warranted.

Error is divided into two classes: random and systematic. Random error results from an imprecision in the ability to measure the variable (including variance beyond the technological ability to measure), unreliability in the measurement (including factors or properties of the variable that remain unknown), or from chance (i.e., a small sample size).

The role of random error in producing the observed results typically can be quantified and used to test hypotheses. In this way, statistical inference

**Table 14: Classification of Random Error**

| | $H_0$ True | $H_0$ False |
|---|---|---|
| Reject $H_0$ | Type I error | |
| Accept $H_0$ | | Type II error |

$H_0$ = hypothesis

can be considered a measure of confidence that either validates or invalidates inferences made by deduction. Such hypotheses are formed so that the null hypotheses ($H_O$) assume a random relationship (unknowable, unpredictable), and the alternative hypotheses ($H_A$) assume a non-random (knowable, predictable) relationship. Random error is known by two types: Type I error and Type II error (Table 14).

Type I error is the probability of rejecting the null hypothesis when the hypothesis is true (i.e., saying that two groups are different when in reality they are not). This error must be avoided or minimized to confidently conclude that a non-random (or statistically significant) relationship exists. The probability of Type 1 error is better known as the p-value. In order to determine the statistical significance of the observation, the p-value is tested again by comparison to the investigator's tolerance for Type I error (commonly designated as the Greek letter alpha—$\alpha$). When the p-value is smaller than alpha, the observation is determined to be "significant" because the likelihood of being "incorrect" is in the range of tolerable. Ideally investigators should be able to determine their own tolerances for Type I error based on the clinical or community consequences for committing the error. More often, the scientific community chooses the tolerance for Type I error by the traditional default value of $\alpha = 0.05$.

"Most of the errors (at least as they relate to statistical inference) center on misuse of the t-test." (3) Repeated use of this or any other method developed for single comparisons allows the tolerance for Type 1 error (alpha—$\alpha$) to compound. Thus, if the alpha is set at 0.05 (the rule of thumb commonly accepted in scientific communities) for one t-test, then after five t-tests the alpha has increased to 0.25. One way to maintain alpha at 0.05 in multiple comparisons is to use methods developed for multiple comparisons (analysis of variance—ANOVA, multiple regression, other multivariate methods, or correction methods such as Bonferroni). Unfortunately, most statistics courses cover the t-test and little else for hypothesis testing. The methods for multiple comparisons are tucked away in the back of the texts and mentioned briefly, if at all, as something for more advanced study.

Type II error is the probability of not rejecting the null hypothesis when the null hypothesis is false (i.e., saying that two groups are similar when in fact they are different) and is usually designated by the Greek letter beta—$\beta$. Type II error becomes important when a clinically important difference is observed between two groups, yet the p-value is non-significant. Statistical power is 1—beta, or stated alternatively, the ability to detect a clinically important difference if one truly exists. The need arises to calculate the statistical power of the study to determine if the study was able to detect a clinically important difference by design.

Statistical power is best addressed during the design of the study, and will usually be adequate if the proper process is followed to estimate an adequate sample size. The general formula for estimating a needed sample size is given by the formula:

$$n = \frac{\sigma^2 (Z_{1-\beta} + Z_{1-\alpha})^2}{(M_0 - M_1)^2}$$

*where:*

n is the sample size.

$\sigma^2$ is the sample variance of the difference between the two groups.

$Z_{1-\beta}$ and $Z_{1-\alpha}$ are the Z-scores (area under the standard normal curve) for the desired statistical power and accepted tolerance for Type I error.

$M_0$ and $M_1$ are the means (expected) in the two groups being compared.

This formula shows the numerator to contain the tolerances for Type I and II errors and the measure of dispersion (variance in this case). Being in the numerator makes the variance and tolerances for Type 1 and 2 errors directly proportional to the sample size requirement. That is, as these values increase, the sample size requirement will increase. As the tolerances for errors decrease, the Z-scores for these tolerances will increase. The denominator includes the magnitude of the observed difference between the two groups squared. Thus, it will take many more participants to statistically validate a small difference between two groups. The target difference used in planning a study should be the minimum clinically significant difference, and the maximum that the investigator would not want to miss.

If the calculated number of participants is used, and all of the other assumptions in the estimate hold true, then a sample size calculation will assure that the study will yield adequate statistical power. Statistical powers of greater than 0.90 will require a larger sample size, and more time, effort, and costs in conducting the study. Statistical powers less than 0.70 can be criticized if there are negative findings because inadequate power is a likely culprit (at 0.50, the statistical power of a true null hypothesis has the same degree of belief as a coin toss). Most clinical research studies are successfully designed with powers between 0.70 and 0.90. The power selected by an investigator should take into consideration the cost of the study as the sample size increases and the confidence or certainty with which the parameters used to calculate the sample size is known. For example, if an investigator suspects that the observed difference between two treatments will be smaller than originally thought, and that the smaller difference will not be clinically important, then the cost of higher statistical power might be considered unwise. Statistical significance of an association is often misinterpreted by investigators and clinicians. Hennekens and Buring (4) outlined four important concepts that, if addressed in the review of statistically significant results, can enhance the likelihood of appropriate interpretation of statistical significance:

- Recall that the p-value should be used as a guide; a p-value, no matter how small, does not exclude chance completely.
- Remember that statistically significant results are not always clinically significant/relevant. The question of *statistical* significance is different from the question of *clinical* significance/relevance. Always look at the magnitude of the difference.
- Use caution in the interpretation of multiple tests of significance. Whenever possible, incorporate appropriate methods to correct or control for this effect.
- Realize that even in the face of statistically significant results (which reflect only on the role of chance), other possible explanations of the significant association between the disease and the exposure should be considered (i.e., systematic errors of bias and confounding described below).

## What Are the Limitations of the Study? The Role of Bias and Confounding

Systematic error affects the internal validity, as opposed to the significance of an experiment. Threats to validity call into question whether or not the study measured what it purports to measure (internal validity) and whether or not the conclusions can be extrapolated to populations beyond the study population (external validity). Saying that a study has internal validity means that it was well conducted. Without high internal validity, external validity has little relevance. Systematic error can generally be divided into biases and confounding, which are briefly discussed below.

*Bias*

Bias is the systematic error in any phase of a study that leads to a result that differs from the truth (5). Because bias is not knowable beyond the abstract concept, it does not easily lend itself to measurement, mathematical adjustments, or corrections to counteract the bias effect. The best protections against bias are those carefully planned into the study design.

Selection bias is one of the most commonly occurring types of bias. Selection bias occurs when the study population is chosen, or selected, so that it is not representative of its target population. This can occur in many ways. Selection bias is most likely to occur in a case-control study where patients are included on the basis of whether or not they have the disease. As stated earlier, in case-control studies, cases should be representative of all persons who develop the disease and controls representative of all persons without the disease. When these conditions are not met, selection bias is likely to be present. For example, if we were to conduct a case-control study of risk factors for myocardial infarction (MI) using patients identified in the outpatient setting (i.e., prevalent cases or MI survivors), bias may be present. Clearly persons who survive an MI are not representative of all persons who develop MIs. Additionally, bias may be present if we limited our case selection to all newly diagnosed cases of MI from a particular hospital. Bias may be present in this situation, as patients from a particular hospital may not be representative of all persons with MI. Finally, hospital-based controls are generally not representative of all persons without disease and therefore using such patients as controls may result in an odds ratio that is biased. Although less common, selection bias can occur in a cohort study if the exposed or nonexposed groups are not representative of the general population that is nonexposed and exposed, respectively. These are but a few examples of the many ways bias can occur in studies. The possibility for similar bias is endless, and the best approach to assess whether selection bias has infiltrated a study is to determine if cases and controls or exposed and nonexposed populations were selected in such a manner as to

be representative of their target populations. The best defense against selection bias is, when possible, to conduct a completely blind randomization scheme using a sample representative of the target population. Under this scenario, participants are recruited, deemed eligible for any group in the study, and then randomly assigned to a group beyond the control of anyone involved in the investigation. Strict compliance with the protocol from randomization through analysis is called an "intention to treat" method. While this furthers the protection against selection bias, analyzing according to intent to treat may reduce the difference between the groups and might drive the need for a larger sample size.

Errors in measurement of exposure or outcome can result in information bias, a broad umbrella term which may also be referred to as misclassification, diagnostic, or calibration bias. These biases result from errors in the categorization of exposure, disease, or both due to imprecise or faulty instruments or methods. Misclassification bias occurs when a participant is coded or classed contrary to his/her true status (e.g., a smoker coded as a nonsmoker). Another way misclassification can occur is with a crossover against protocol. This occurs in randomized controlled trials when a participant undertakes self-administration of the other group's treatment or a treatment that at least mimics that of the other group. This usually goes unreported and may reduce the difference between groups. Misclassification errors throughout the data collection process are best prevented with vigilant protocol compliance and monitoring with quality audits of the data to minimize inaccuracies in data collection. Information biases are errors in the data collection or abstraction. This type of bias includes recall bias (cases remembering previous exposures differentially compared to controls) and interviewer bias (interviewers asking cases and controls differently about previous exposures). Also, nonresponse, missing, and unknown data constitute information bias. This bias affects the accuracy of the analysis and the confidence of the conclusion. Strict protocol compliance and double blinding/masking (i.e., keeping participants and study staff unaware of treatment assignment), when feasible, are the best methods for minimizing information bias.

Diagnostic or detection bias is the differential assessment of outcome between the exposed and unexposed groups. This type of bias is most common in cohort studies where exposed and nonexposed persons are followed over time. For example, this bias would be present if an investigator assessed disease/outcome status differently in the exposed (e.g., smoker) and unexposed groups. The differential monitoring of participants may be quite subtle. It is for this reason that study staff that assess outcomes should be masked/blinded to the exposure of study participants.

Calibration bias reflects inaccuracies in measurements either inherent in the device or due to the operator. This type of bias is especially relevant in technology assessment, where statistical tests that are sensitive to random error might conclude that two technologies are correlated, although any existing systematic calibration error might lead to differences in clinical decisions resulting from the use of the two technologies. Calibration bias can be quantified, but only after it is discovered and addressed.

Publication bias is a failure to publish solely on the basis of the direction or magnitude of difference in the study results. The direction of the result refers to whether or not the finding concurs with popular belief. Especially in situations where the popular belief was established with small samples, the "tug-of-war" from a divergent finding should be a part of the published debate in the scientific literature. The magnitude of difference in publication bias usually does not work against findings of large magnitude; it is typically findings of small magnitude that are discriminated against. While it is true that many of these studies could simply be underpowered, it is not a rule. The scientific community is conditioned to want significant findings (small p-values), so negative findings are not deemed interesting. This could lead the medical community astray if two or three significant studies on a topic are published, and dozens of negative findings on the same hypothesis are not published. If the preponderance of the evidence were known, the readership might not be so confident in rarely found significant results. The solution cannot be to publish the results from every permutation of hypotheses; there would never be enough space to gather that much information. The negative effects of publication bias can be minimized through participation in scientific meetings and societies, in abstract and poster forums, and in critical review of what does get published in the medical literature.

### *Confounding*

Confounding is the second general category of systematic error. Confounding occurs when a third variable comes into play that is related to both the exposure and outcome variable. The confounding variable affects the character of the relationship between the exposure and outcome. The effect can be to make the relationship appear larger or smaller, or to

completely mask the relationship. There are some approaches to handling the issue of confounding. One approach is to match individual or group data on the variable suspected of being the confounder. Another approach is to utilize stratification or adjustment techniques in the analysis of the data. Sometimes the relationship between an exposure and outcome is dependent on the level of a third variable. This is called interaction, because two exposures are interacting to alter the risk of disease.

## Are the Study Results Relevant to My Patients?

Once we conclude that a study maintains internal validity, we turn our attention to whether the study results are applicable to our patients. Many epidemiological studies and clinical trials are conducted in highly select populations. For example, in the Prolyse in Acute Cerebral Thromboembolism (PROACT) II trial, patients with acute ischemic stroke of less than 6 hours in duration were randomly assigned to an intra-arterial thrombolysis group or a control group (6). However, patients with symptoms for greater than 6 hours were excluded. As noted by Gross and colleagues, "The implications for clinicians caring for patients in the 'real world' are that most of their patients with severe strokes may differ from trial enrollees in important ways and that they should be extremely cautious when selecting candidates for thrombolysis." (5) Nonetheless, the results of this trial were valid for patients similar to those included in the study.

## Conclusion

Statistical and epidemiological principles, measures, and tests are important tools that can be used to facilitate clinical decision-making and are key in implementing and evaluating clinical research. This two-part series provides an overview of basic concepts that can enhance a physician's review of the medical literature or planning/evaluation of a study. Understanding the purpose, advantages, and limitations of various study designs, the appropriate use of tests and statistical significance, and the study limitations is key to the appropriate use of scientific information in clinical research and clinical practice.

## REFERENCES


1. Krousel-Wood MA, Chambers R, Muntner P. Clinician's guide to statistics for medical practice and research: Part I. Ochsner Journal 2006; 6(2):68-83.
2. Willenheimer R. Statistical significance versus clinical relevance in cardiovascular medicine. Prog Cardiovasc Dis 2001;44(3):155-167.
3. Glantz SA. Primer of Biostatistics, 5th edition. New York: McGraw-Hill, 2002.
4. Hennekens CH, Buring JE. Epidemiology in Medicine. Boston: Little, Brown and Company, 1987.
5. Sackett DL. Bias in analytic research. J Chronic Dis 1979;32(1-2):51-63.
6. Gross CP, Mallory R, Heiat A, Krumholz HM. Reporting the recruitment process in clinical trials: who are these patients and how did they get there? Ann Intern Med 2002; 137(1):10-16.