Assessing the Medical Literature: Let the Buyer Beware

Victor A. Ferraris, MD, PhD, and Suellen P. Ferraris, PhD

From the Division of Cardiovascular and Thoracic Surgery, University of Kentucky Chandler Medical Center, Lexington, Kentucky

As many as 30% of journal articles may contain errors. Most of these errors involve the use of simple statistical tests or elementary principles of research design. Assessment of the thoracic surgical literature involves cautious circumspection. This does not mean that it is necessary to have in-depth knowledge of sophisticated statistics, rather it means that common sense understanding of a few principles of research design and simple statistics are necessary to determine the usefulness and believability of literature publications.

(Ann Thorac Surg 2003;76:4–11)

In reviewing an article that presents the results of a new intervention or innovative experiment, the reader needs to answer at least three questions: (1) Are positive findings really positive? (2) Are negative results really negative? and (3) Is there any evidence of experimental bias?

Positive findings in a research report or clinical trial need to contain some assessment of how the authors have inferred that the findings are positive in favor of one treatment or intervention. This requires knowledge of simple statistics and awareness that a competent biostatistician must review statistics that are more complex. Most errors in the medical literature occur with the use of the simplest statistics, including the Student's *t* test, χ^2 test, and analysis of variance. Paradoxically, fewer errors occur with complex statistics because biostatisticians are usually involved in the study design and analysis.

Negative findings need to be backed up with some assessment of statistical power and estimates of the type II statistical error. Falsely negative results occur with too small of a sample size in the study group. A simple understanding of the principles of experimental design and realizing how sample size affects the type II error is the key to believing any negative findings presented in the literature.

Experimental bias is omnipresent. Important maneuvers such as using control groups, randomization, or statistical adjustment techniques are essential in reducing bias in both prospective and observational studies. When some effort at bias reduction is not used, then research results are suspect. Authors have to convince the reader that they did the best they could to reduce bias.

Understanding these basics of experimental design and simple statistics allows the surgeon to tame the thoracic literature and gain useful insights that help improve the practice of thoracic surgery.

Why Read the Thoracic Literature?

At least three of the reasons for thoracic surgeons to read the medical literature are to (1) improve patient care by self-education, (2) learn about research and cutting-edge technology, and (3) educate peers and students about clinical care. It is axiomatic that reading the literature requires some understanding of statistics. The aim of this review is to simplify and direct the reader's understanding to the key and essential features of a literature article, identifying common pitfalls along the way.

What to Expect From the Thoracic Literature

What can a thoracic surgeon expect to get from the literature? How reliable is the information published in these articles? One rather extreme view is offered by Williamson and coworkers [1] who reviewed 28 articles assessing the quality of the medical literature. These authors summarized their findings as follows: "The average practitioner will find relatively few journal articles that are scientifically sound in terms of reporting usable data and providing even moderately strong support for their inferences" [1]. Stan Glantz [2] presents a slightly more optimistic view of the cardiology literature. He found that as many as 27% of a random sample of articles from a cardiology journal contained errors in the use of simple statistics. Almost every assessment of the medical literature has found significant errors in research design or use of statistics. No wonder why we have heard thoracic surgeons say that statistics are meaningless, too difficult to understand, or are used to show whatever you believe. Does this mean that we cannot trust or believe the medical literature? Of course it does not. It does suggest that surgeons have to be cautious about interpreting the thoracic literature, just like any other aspect

Address reprint requests to Dr Ferraris, Division of Cardiovascular and Thoracic Surgery, C208, Chandler Medical Center, University of Kentucky, 800 Rose St, Lexington, KY 40536; e-mail: vferr2@uky.edu.

Published by Elsevier Inc

0003-4975/03/$30.00
PII S0003-4975(03)00430-2

Table 1. Three Questions to Ask About a Literature Article

Question	Knowledge Required
1. Are positive results really positive?	Requires understanding of simple statistics (χ^2 test, *t*-test, and analysis of variance) and understanding of *p* values (type I error)
2. Are negative results really negative?	Requires understanding of relationship between sample size and statistical power (type II error)
3. Is there any evidence of experimental bias?	Requires understanding experimental design and bias reduction measures (stratification, randomization, controls, and so forth)

of their surgical practice. The attitude of letting the buyer beware seems to apply.

What Statistics Do I Have to Know?

An understanding of literature articles is not about creating in-depth knowledge of sophisticated statistical methods that appear in the literature. In fact, it is the opposite. The literature articles that contain the least errors in research design and statistical methods are those with the most complex statistical methodology. This is true because almost all of these complex articles involve active participation by experienced biostatisticians, often as coauthors. Literature articles that require understanding of complex statistics are almost always accepted on blind faith, probably as they should be, because of this involvement by experienced biostatisticians. Ironically the errors in the literature usually involve simple mistakes, such as improper research design or misuse of elementary tests of hypotheses.

In order to get the maximum benefit from thoracic literature, it is necessary to know key features about the basic tests of hypothesis, such as the χ^2 test and the Student's *t* test. An understanding of the simplest hypothesis testing and of the basic principles of research design and sample size is more than enough to provide the basis for critical thinking about thoracic literature. It is useful to summarize these principles in the form of three questions that should be asked about every original literature article that compares two or more treatments or presents a new intervention (Table 1). By answering these three questions, the surgeon becomes a critical thinker and uses the thoracic literature with discrimination rather than with blind faith. The ultimate benefit of this exercise is to improve patient care, physician education, and the quality of thoracic literature.

Question 1: Are Positive Results Really Positive?

Most literature articles present some new information about a treatment or intervention. It is easier to get an article published if there are positive findings presented, which is a phenomena known as publication bias. Many negative articles are not deemed publishable either by authors or by editors. Therefore the critical reader of the literature starts out with a biased sample of articles to choose from, mostly those containing positive results. As a corollary, most errors in the medical literature occur in articles that contain positive findings.

Describing the Variables

Karl Pearson, in the early 1900s, was one of the first to realize the inherent lack of precision of natural measurements. He expanded on the works of others, such as Gauss and Bernoulli, by describing measurements in terms of a distribution, rather than as an exact value. In Pearson's terminology, the distribution of numbers could be written as a mathematical formula that provides the probability that an observed number will have a given value. What value the number actually takes in a specific experiment is unpredictable. Pearson only talked about probabilities of values and not about certainties of values. It was not until this fundamental understanding of the randomness of observations was appreciated that modern statistics could proceed [3]. Pearson described four factors that completely explain a measurement. These four factors are shown in Table 2 and are the basis for most of the statistical tests used to compare continuous variables (eg, blood pressure, age, ischemic cross-clamp time). Calculation of these factors is a relatively simple matter thanks to the availability and power of modern computers. The values of Pearson's factors appear in any

Table 2. Pearson's Parameters That Describe the Measurement of a Variable

Parameter	Definition	Mathmatical Definition
Mean	The central value about which the measurements scatter	$X = \Sigma X_1/n$
Standard deviation (equal to the square root of the variance)	How far most of the measurements scatter about the mean	$s = \sqrt{s^2} = \sqrt{\Sigma(X_1 - X)^2/n}$
Symmetry	The degree to which the measurements pile up on only one side of the mean	
Kurtosis	How far rare measurements scatter from the mean	

s = standard deviation; s^2 = variance; χ = mean of measured value; χ_i = measured values of a variable; n = number of measurements.

one of a myriad of statistical software programs and form the basis of simple and complex statistical calculations.

What Does the t *Test Really Do?*

One of the most helpful and widely used tests to measure a positive difference between two treatment groups or interventions is the Student's *t* test. The Student's *t* test was originally described by William Sealy Gosset in 1904. He was forced to publish his observations using a pseudonym (ie, Student) because of limitations imposed by his employer, the Guiness Brewing Company of Dublin. Gosset made critical comments about comparing small groups of observations, and he realized that small samples could be described by a distribution called the *t* distribution or *t* value:

$$t = \frac{(\bar{x} - \mu)}{s/\sqrt{n}}, \qquad (1)$$

where $\bar{x}$ = mean of randomly sampled values of a parameter (ie, the mean of a small sample), μ = the true population mean value, s = the sample standard deviation, and n = the number of observations or the sample size.

Largely by trial and error, Gossett noted that the value $x_i - \mu$ (called the sampling distribution of the mean) is also normally distributed (ie, the *t* distribution from Equation 1 is a series of values distributed about the mean like the "Gaussian" normal distribution). The Student's *t* test is used to test hypotheses about differences between two (and only two) population means. It makes use of the symmetry of the *t* distribution and compares the *t* statistic calculated from each of the two sample *t* distributions with critical values. By applying a little algebra to Equation 1, the *t* statistic for two randomly obtained sample means is (again the computer does this for you):

$$t\,\text{statistic} = \frac{(X_1 - X_2) - (\mu_1 - \mu_2)}{s_p\,\sqrt{(1/n_1) + (1/n_2)}}, \qquad (2)$$

where the subscripts represent the values for sample 1 and sample 2. The "null hypothesis" suggests that $\mu_1 - \mu_2$ is zero (ie, there is no difference between the true population means) and this allows the computer to calculate the *t* statistic and compare it with a critical value. If the calculated *t* statistic is greater than the critical value, then the null hypothesis is rejected and the two samples are significantly different. The astute reader, both of this article and of the literature, quickly realizes that the forgoing discussion of the *t* test gives the principles behind the *t* test, but exposes the computer as the workhorse that does the calculations.

The *t* test has stood the test of time and arguably, has proven to be more useful than any other single statistical test. There are three fundamental assumptions implicit in *t* test comparisons: (1) The variances (and hence the standard deviations) of each sample are nearly equal. (2) The populations from which the samples are drawn are normally distributed. (3) Each sample is independent of the other (ie, samples are drawn from independent populations). Over the years, statisticians realized just how reliable the *t* distribution is. In fact, it is likely that some or all of the above assumptions are not necessary in order for the *t* test to give a very accurate approximation of the comparison of two population means.

Table 3. What Does the t *Test Really Do?*

1. Compares means between 2 groups and only 2 groups!
2. Measures the degree of difference between groups (the larger the *t* statistic, the more likely that a difference exists between the two groups).
3. Allows calculation of probability that a difference exists between the groups (the famous *p* value).

It is surprising how often the *t* test is used improperly in the medical literature [2]. Most of the errors arise from two critical areas. First, the basic assumptions of the sample population are not met, and second, the *t* test is used for multiple comparisons (ie, comparing more than two groups or using the *t* test multiple times for different comparisons) (Table 3). It is not necessary to know how to calculate the *t* distribution, because computers do it for you. It is necessary to know the inherent assumptions and how the *t* test is used. This is more important than any of the calculations used to derive the *t* test results.

If an article in the literature uses the *t* test to compare multiple groups, the likelihood of finding a spuriously significant difference between one or more groups is high. As an example, suppose that a *t* test comparison of two means of 5 independently sampled groups (one of which is a control group) reveals a ratio that is more than the critical value for the 5% level (ie, there is a 5% chance that the means are not significantly different, which is the type I error or the α-error). If the *t* test is used to compare the control group with other means in the group, then the probability that two test results will be significant is 0.05 × 0.05 = 0.0025. The probability that neither test will be significant is 0.95 × 0.95 = 0.9025. The probability that at least one of the two test results will be significant is 1 − 0.9025, or 0.0975. Therefore, the probability of incorrectly deciding that the members of either one or both pairs of means are significantly different using two tests is nearly twice the probability of making the same error for a single test (0.0975 vs 0.05). In general, the probability of finding at least one spuriously significant independent result is calculated as follows:

$$P = 1 - (1 - \alpha)^n, \qquad (3)$$

where P = probability of at least one spuriously significant result, α = the original type I error for a single test, usually 0.05, and n = the number of times that the *t* test is used for multiple comparisons. Thus using the *t* test to compare a control group mean with four other group means results in a 19% (ie, 1 − $[0.95]^4$ or 0.19) chance that at least one of the comparisons is deemed different but really is not. If you wanted to compare every combination between the 5 groups, there would be 24 tests and a 71% chance that at least one of the comparisons is a false positive! Beware of positive results that use multiple *t* test comparisons.

Table 4. Tests of Multiple Comparisons Used in Conjunction With Analysis of Variance

Multiple Comparison Test	Ease of Use	Most Severe
Bonferroni	++++	−
Dunnett	+	+
Tukey	++	+
Scheffe	−	+++

What About Multiple Comparisons?

Often literature articles present the results of interventions or observations in more than two groups. Investigators comparing three or more group means frequently examine each possible pair of groups separately, using the *t* test to examine each pair [4]. This is not correct! In testing multiple groups by pairs, problems arise because of the number of tests made and because of the lack of independence. The correct way to analyze different treatments or measurements in more than two groups is by means of analysis of variance (ANOVA) combined with statistical tests specifically designed for multiple comparisons between group means. Analysis of variance indicates whether there are differences among the population means of the groups being compared, but it does not indicate which groups, if any, differ from the others. Analysis of variance generalizes the *t* test from two groups to three or more groups. It replaces multiple *t* tests with a single F test (named after Sir Ronald Fisher), which assumes that the underlying group means are all equal (again, the null hypothesis, this time for ANOVA). Analysis of variance is able to generate a single test statistic by analyzing variance (ie, standard deviation) within each group and between different groups. It is not necessary to know the computational methods used to determine these group variances, because the computer does it. The assumptions implicit in ANOVA and the methods used to compare group means when more than 2 groups are being studied are the key and essential features that the critical reader needs to know.

In order to compare individual group means in three or more groups, it is necessary to use multiple comparison techniques. One simple way to look at these techniques is to think of them as corrected *t* tests with more stringent critical values that account for the lack of independent sampling and the need for multiple comparisons. Perhaps the simplest multiple comparison test (and the most stringent) is the Bonferroni test (Table 4). Using this test, the critical probability of the *t* test is divided by the number of group means in the multiple comparison groups. For five group means the *p* value for a statistically significant comparison is 0.05/5 or 0.01. That implies that for any two means to be significantly different at the 5% level, the corrected *p* value for a *t* test comparison must be less than or equal to 0.01.

The names of these multiple comparison tests are almost always included in the Methods section of a literature article. They contain names like Scheffe, Tukey, Bonferonni, Dunnett, Newman-Keuls, and Duncan. Most statistical software packages will spit out all of the multiple comparison tests as a matter of routine in ANOVA. What is important is that the authors have used the multiple comparison tests, not necessarily which one was used (Table 4). A special case of ANOVA occurs when measurements of group means are made at repeated intervals. For example, blood pressure measurements in response to antihypertensive medication can be made every week for a month. In order to see if the drug caused a decrease in blood pressure, ANOVA with repeated measures should be used. Again, computer software can easily perform the necessary calculations and it is necessary to use some test of multiple comparison more stringent than the *t* test to evaluate any differences between the different time points of measurement.

Godfrey [4] reviewed articles in volume 321 of the New England Journal of Medicine that compared interventions or measurements in three or more groups. Of the 50 articles he examined, only 27 (46%) used appropriate ANOVA with multiple comparisons. The majority of the articles used multiple *t* tests to compare group means. Multiple *t* tests for group comparisons introduce a large and misleading increase in the probability of finding at least one significant test result where no real population difference exists. The study by Godfrey was done in 1989. Undoubtedly, things have changed since then, but the best advice to the average thoracic surgeon is to be cautious about literature articles that report multiple group comparisons without using ANOVA and tests of multiple comparisons.

What about Discrete Variables?

Many of the measurements that are presented in the thoracic literature contain *yes* or *no* variables or variables that fall into categories, not continuous ranges. These variables are categorical, because there is no arithmetic relationship between the different classifications. There are actually two types of categorical variables, nominal and ordinal. Nominal variables fit into categories and each category is not related to the other categories by a mathematical relationship. Nominal variables describe a quality of a person or thing and nominal data are often described in terms of percentages or proportions. Examples include survival or death from an operation, cure or recurrence from cancer treatment, the presence or absence of postoperative stroke, or patient race (white, Asian, African-American, Latino, and so forth). If there is an inherent order among the categories, then the observations are ordinal variables. Ordinal variables measure the amount of risk a patient has or the type of graded therapy that is appropriate. For example, lung cancers are staged into ordinal variables (eg, stages I–IV for nonsmall cell lung cancer). The grades of Papanicolaou smears are examples of ordinal variables. It is important to make the distinction between nominal and ordinal variables, because the statistical methods used to draw inferences about these two types of categorical data are different and should not be confused or intermingled.

Categorical data composed of either nominal or ordinal variables are often presented in contingency tables

Table 5. Contingency Table for the Relationship of Hormone Replacement Therapy and Mortality in Postmenopausal Women[a]

	Number of People		
	Deceased	Alive	Totals
Current HRT use	574 (a)	8,483 (b)	9,057 (r_1)
No HRT	2,051 (c)	17,520 (d)	19,571 (r_2)
Totals	2,625 (s_1)	26,003 (s_2)	28,628 (N)

[a] Letters in parentheses are notations used in describing parameters of contingency tables in the text [6]

HRT = hormone replacement therapy.

that show the number of patients belonging to each category. As many as 70% of literature articles contain some form of contingency table [5]. Contingency tables always deal with counts or classification of discrete variables. These tables both analyze categorical data for significant trends and describe characteristics of patients under study.

An example of the use of contingency tables is found in an article published about the relationship of hormone replacement therapy (HRT) to mortality in the Nurses' Health Study Cohort [6]. Table 5 is a summary of the mortality data dealing with a cohort study of 38,261 nurses, 18,690 of whom took HRT after menopause. An obvious question (ie, hypothesis) is whether HRT is associated with increased mortality in this cohort. Table 5 is the simplest form of a contingency table used to represent this data. Importantly there are at least three different methods of sampling that give rise to the frequencies set out in contingency tables: (1) cross-sectional sampling where a total group of subjects (cross-section) is divided into various categories, (2) selection and study of a predetermined group who possess a given characteristic compared with a control group who do not possess the characteristic (so-called case-control or cohort study), and (3) randomized clinical trials in which two samples of predetermined size are constituted at random. All of these methods of sampling are analyzed in the same way [7]. Table 5 represents an example of using a contingency table to summarize a case-control study excerpted from the Nurses' Health Study [6].

A contingency table, such as the one shown in Table 5, allows a calculation of the odds ratio, or the odds that a woman who takes HRT will die compared with a woman who does not take HRT. The odds from a contingency table are the number of patients who had an event divided by the number of patients who did not have the event. From the information shown in Table 5, the odds of death for a woman who takes HRT is 574 out of 8483, or 0.067. The odds ratio is the cross product of the 2 × 2 contingency table. The letters in parentheses in Table 5 generalize the various calculations from the contingency table. For example, the odds ratio is the ratio of mortality in HRT users compared with nonusers:

$$\text{Odds ratio} = (a/b)/(c/d) = ad/bc$$

$$= (574{*}17{,}520)/(8{,}483{*}2{,}051) = 0.58. \tag{4}$$

The odds ratio should be distinguished from the relative risk. The relative risk is the ratio of two probabilities, again obtained from the contingency table. Probabilities are the number of patients who had an event divided by the total number of patients. From Table 5, the relative risk of death for a current HRT user is:

$$\text{Relative risk} = \frac{\dfrac{\text{Number in group 1 with trait}}{\text{Total in group 1}}}{\dfrac{\text{Number in group 2 with trait}}{\text{Total in group 2}}}$$

$$= \frac{574/2625}{8483/26003} = 0.67. \tag{5}$$

For low risk events, the odds ratio and the relative risk are close but not equal, with the odds ratio always being bigger than the relative risk [8]. Many journal articles use the odds ratio because they are easily obtained from case-control studies and from logistic regression. For example, the adjusted odds ratio (ie, adjusted for multivariate regression) is the exponent of the logistic regression coefficient.

This means that HRT users are 0.67 times as likely to die as are non-HRT users. To use statistically proper terms, HRT use is associated with a decreased risk of death compared with nonuse in the study population. Often the odds ratio associated with a contingency table contains a confidence interval. If the confidence interval encompasses unity, then the comparison between the interventions or treatments in the table is not significantly different.

It is ironic that the two most commonly used tests for comparing rates and proportions of discrete variables are named after arch rivals (almost arch enemies) in the statistics world, Karl Pearson and Ronald Fisher [3]. Fischer's exact test and the Pearson χ^2 test are the most commonly used tests of hypotheses generated from 2 × 2 contingency tables. These statistical tests have some unique features that make them ideal for testing medical hypotheses. First, the χ^2 test is used to test hypotheses about two groups or more than two groups. Second, the χ^2 test is a nonparametric test. This statistical term implies that the χ^2 test can be used in populations that are not normally (randomly) distributed. This is noticeably different from the *t* test. In the example shown in Table 5, the Pearson χ^2 test is used to test the hypothesis that HRT users are less likely to die than nonusers. To use the χ^2 test it is first necessary to estimate the expected mortality rate by summing the rows (row 1 and row 2) and columns (column 1 and column 2) and dividing by the grand total (n). The expected rate of mortality (E) in HRT users is compared with the observed mortality rate

(O) according to the following relationship originally described by Pearson:

$$\text{chi-square } (\chi^2) = \sum_{n=2} (O - E)^2 \div E. \qquad (6)$$

Again, as with the *t* test, values for the test statistic, χ^2, can be compared with a range of values (called the χ^2 distribution) by a computer program with resultant comparisons made between the critical value for a predetermined type 1 error (usually 0.05 or less). In most cases it is not even necessary to calculate the expected values (E), because the computer will do this for you. Incidentally, the χ^2 test applied to data in Table 5 suggests an association between HRT use and reduced risk of death ($p < 0.05$ by the Pearson χ^2 test) [6].

A special case arises when one of the values in the 2 × 2 table is small (ie, less than 5 units). When the expected frequency of one of the cells of a contingency table is smaller than 5, it is possible to simply list all the possible arrangements of the observations and then compute the exact probabilities of each possible arrangement of the data. This correction is called Fischer's exact test. The most common trap with this is that cells in the 2 × 2 table that contain small numbers need to use Fischer's exact test to compute the test statistic, and this cannot be done automatically by the computer [9]. Using the χ^2 test when one or more of the cell contents in a 2 × 2 contingency table are small (ie, failure to use Fischer's exact test appropriately) can result in spuriously significant critical χ^2 values and incorrect inference from the contingency table.

The χ^2 test can be expanded to contingency tables that include more than 2 columns or rows. When the contingency table includes more than 2 rows or columns, then a more general formula is used to calculate the expected values. This calculation of the expected values depends on the degrees of freedom of the N x M contingency table. It is not too important to know how to compute the degrees of freedom or the expected values, because the computer usually does this with great accuracy.

When χ^2 analysis is used to analyze contingency tables with more than 2 rows or columns, the situation is analogous to the multiple-comparison procedures used in ANOVA. Multiple comparisons between the various cells in a contingency table with more than 2 rows or columns needs to have a correction applied to the *p* value in order to infer significant differences between the compared cells. Again, the important thing to recognize is that a corrected *p* value needs to be calculated, not how to do it. If the authors of a journal article do not realize that corrected *p* values are necessary in multiple comparisons, then the results must be suspect.

What About Ordered Categories?

Journal articles often present outcome data in the form of ordinal variables, again using contingency tables. Examples include patients who are worse, unchanged, or better after an operation or therapy. Data that can be qualitatively ordered should be analyzed using the Wilcoxon-Mann-Whitney *U* test or tests that rely on ranking the outcome measurements. These statistical tests compare the average ranks of two or more samples by comparing a calculated probability with a critical value of the relevant rank-sum test. Names like Wilcoxon, Mann-Whitney or Kruskal-Wallace are the usual rank-sum tests used by most statistical computer packages. Again, the computer does the work. It is not so important to know how to do the calculations. However, it is important to know when rank-sum tests are necessary to analyze ordinal data. Journal articles that contain either ordered input data or outcome data with qualitative ranks need to be scrutinized carefully to assure that the authors used a rank-sum test to analyze the results. Using other statistical tests, like χ^2, to analyze this type of ordered data wastes information and produces results that are insensitive to changes in the various orders [10].

Table 6. Relationship Between Preoperative Template Bleeding Time and Postoperative Blood Loss [11]

Template Bleeding Time	Number of Patients	Average Chest Tube Drainage per Patient
Bleeding time ≤ 4 minutes	5	423 ± 258
Bleeding time ≥ 8 minutes	5	745 ± 252[a]

[a] $p > 0.05$.

Question 2: Are Negative Results Really Negative?

Should a costly drug be used to limit blood loss after operation? A well-done study to answer this question can provide valuable information to clinicians, whether the results are negative or positive. As previously mentioned, there is a publication bias in favor of positive findings in the medical literature. Nonetheless, negative results that answer questions about drugs or interventions are every bit as helpful as positive results.

When negative studies appear in the literature, an important question needs to be answered. If the therapy or intervention has no effect, were enough subjects studied to confidently assert that there really is no difference between the intervention and the control? Ideally the investigators should ask (and answer) this question before starting the study. Unfortunately, we live in a less than ideal world.

Table 6 is an example of a negative study that appeared in the thoracic literature [11]. This study evaluated the blood loss in coronary artery bypass graft patients who had template bleeding times measured before operation. The authors asked the question: "Is a prolonged template bleeding time associated with increased postoperative blood transfusion?" Based on results such as those shown in Table 6, they concluded that there was a trend toward increased blood loss in the high bleeding time group, but that this trend was not statistically significant. The authors then concluded that preoperative bleeding time is not useful in detecting patients who will need increased blood transfusion after an operation. How certain can the reader be about this conclusion? There were only 5 patients in the high bleeding time group. Is this enough to conclude with confidence that bleeding time is not helpful in this setting? To answer this question the reader needs to know a few things about type II statistical errors.

When negative findings appear in the literature, the relevant statistics have nothing to do with the conventional tests of inference that we have discussed up to this point. In order to evaluate and understand negative findings, it is necessary to know about statistical power and type II errors. The ability to detect a positive effect depends on the size of the treatment effect, the variability (ie, standard deviation or variance) within the study population and the size of the sample. Bigger samples make it easier to detect an effect and smaller samples make it harder. Authors use tests of statistical inference to try to find a significant treatment difference; when they do not, they often conclude that there is no effect of the intervention. In reality, all they have done is failed to find an effect from their intervention. This does not prove that there is no effect. The difference is subtle but important.

From the point of view of the thoracic surgeon reading the literature, studies that fail to show a significant treatment effect may lack the power to detect the effect because the sample is too small. Statistical tests of inference like the *t* test and the χ^2 test tell you how likely it is that two or more populations are different (the type I error or the chance that the authors have uncovered a false positive finding). They do not tell you how likely it is that there really is a treatment effect, but it was missed because the sample was too small (type II error or the false negative rate). Other statistical considerations are necessary to determine the type II error. The probability of a type II error is not one single value like the probability of a type I error. There are multiple values of the type II error depending on the sample size studied and the magnitude of the treatment effect.

In an ideal world, authors would tell the reader how reliable the negative results are from their study. However, it is rare that any mention of the type II error or its complement, the statistical power, is ever mentioned in a journal article. The type II error is a probability of asserting a negative finding when there is really a positive treatment effect. For example, the type II error for the negative findings displayed in Table 6 is less than 0.5. These values are usually obtained from data tables based on sample size, the magnitude of the treatment effect (ie, the difference between the mean values of the chest tube blood loss in the low bleeding time group compared with the high bleeding time group), and the standard deviation of the group mean. The type II error is often summarized as a percentage called the statistical power, defined as 100 x (1 minus the type II error). Acceptable limits for the type II error (often called the beta error or β) are usually more lenient that those for the type I error (alpha error), which is typically 0.1 for the beta error versus 0.05 for the alpha error. The statistical power for the negative observation depicted in Table 6 is approximately 50%. Another way to state this is that about 13 to 15 patients in each of the two groups (low and high bleeding time groups) are necessary in order to state that with 90% statistical power there is really no difference between the groups. With only 5 patients per group, there is as much as a 50% chance (such as flipping a coin) that the negative observation shown in Table 6 really is a positive finding, but the sample size studied was too small to detect a difference. Therefore, would you say that prolonged bleeding time is not associated with increased postoperative blood loss? Based on results such as those in Table 6, we would not. Beware of the negative study with small samples.

Coincidentally most statistical software does not compute the type II error. Perhaps this is why journal articles rarely contain information about type II errors and statistical power. Nevertheless another important point about type II errors needs emphasis. Investigators who are contemplating a trial or research project should determine what sample size needs to be entered into the study before undertaking the trial. Estimates of the treatment effect can be used to judge the necessary sample size required to obtain a type I error of less than or equal to 0.05 with a type II error of 0.1 or less. The required sample size can be computed from statistical software or looked up in statistical tables. Most large clinical trials benefit from the involvement of an experienced biostatistician who does these calculations before starting the trial. Lacking biostatistician involvement, small trials with fewer patients often do not adequately considered the statistical power and sample size relationships before beginning the study. These small negative trials can confuse and confound the unwary reader of the literature.

Question 3: Is There Any Evidence of Experimental Bias?

Perhaps the most difficult question to answer about literature articles is whether there is unrecognized or unstudied bias in the results presented. Bias implies a predilection or prejudice that slants the results in favor of one treatment over another. The careful reader must be on the lookout for bias in any comparative study. The conscientious author will try to reduce bias as much as possible by designing experiments with adequate checks and balances. However, it is axiomatic that bias exists in all studies and that even the best, most careful study design can only minimize, but not eradicate, inherent bias.

Use of nonrandomized treatment groups is common in the thoracic literature and must be differentiated from articles that report comparisons between two or more groups. Why is it that not all studies use randomized treatment groups or controls? The simple answer is that randomized treatments or use of control groups is not possible in many cases. Some studies are just descriptive and, for one reason or another, do not need a control group. An example of this type of literature article would be a case report that describes an unusual circumstance or operation in one or few patients. In some cases, "uncontrolled" literature reports account for significant advances. In other cases, a reader can get unbiased estimates of population variables for a single group without the use of any control group.

How does the careful investigator reduce bias in studies that compare treatment groups? There are two ways that will do the trick: (1) statistical methods and (2) experimental designs. Statistical methods to reduce bias include determination of inter-observer and intraobserver variability, use of propensity matching [12], ANOVA, and multivariate regression modeling. Experimental designs to reduce bias include use of control subjects, randomization, blinding of investigators or patients, or both, and patient stratification. Despite the use of any or all of these methods, bias still exists. The randomized, controlled trial is the purist example of a study design to reduce experimental bias. Carefully designed, prospective, nonrandomized studies with statistical regression adjustment are an example of the use of statistics to reduce bias in the interpretation of treatment effects in two or more groups.

There have been some rather dramatic examples of investigators (and journal readers) who were befuddled because of lack of adequate bias reduction. In the late 1950s, Mitchell and coworkers [13] advocated ligation of the internal mammary artery (IMA) for relief of angina. The operation to ligate the IMA was rather simple, which was a cut down in the upper chest along the sternal borders. These authors reported dramatic results with relief of angina in as many as 60% of patients. It was only when a randomized, blinded trial was conducted to test this treatment effect that it was obvious that there was no benefit from IMA ligation. This was the beginning of the understanding of the placebo effect. In 1959, Cobb and coworkers [14] conducted a blinded study that probably could not be done today. They made an incision on the upper chest of some patients with angina but did not ligate the IMA and did not tell the patients whether they had IMA ligation or not. Patients who did not have their IMA ligated were just as likely to have relief of their angina as were those who did. The study by Cobb and coworkers [14] is one of the earliest studies to identify what is now commonly accepted as the placebo effect. Angina pectoris is notorious for responding to placebo in as many as 60% of patients. The placebo effect has confounded many drug treatments for angina. Beware of trials and interventions without adequate bias reduction. How many times have articles appeared in the thoracic literature espousing a new heart valve or a new procedure without adequate controls [15], only to have the device or treatment recalled or removed from the market after use in many patients? [16] Lack of adequate bias reduction is both a trap for the reader of the thoracic literature and a source of harm to patients. To avoid falling into this trap, the conscientious reader (and author) needs to understand simple principles of experimental design. It is not necessary to know how to use complex statistical maneuvers to reduce bias. It is necessary to know when these methods should be used.

Comment

In conclusion, many literature articles are not scientifically complete. Common literature errors include misuse of simple statistics, failure to account for type II errors, and incomplete experimental design with failure to account for statistical bias. The important message in reading literature articles is to: (1) be certain of positive results, (2) be skeptical of negative results, and (3) assume experimental bias exists.

The authors wish to thank Dr Gary Grunkemeier for his insightful and invaluable comments. His efforts polished the work of amateurs and added a professional touch.

References

1. Williamson JW, Goldschmidt PG, Colton T. The quality of medical literature: an analysis of validation assessments. In: Bailar JC, Mosteller F, eds. Medical uses of statistics, 1st ed. Waltham, MA: NEJM Books, 1986:370–91.
2. Glantz SA. It is all in the numbers. J Am Coll Cardiol 1993;21:835–7.
3. Salsburg D. The lady tasting tea—how statistics revolutionized science in the twentieth century. New York: A.W.H. Freeman/Owl Book, Henry Holt and Company, LLC, 2002: 1–340.
4. Godfrey K. Comparing the means of several groups. In: Bailar JC, Mosteller F, eds. Medical uses of statistics, 2nd ed. Boston, MA: NEJM Books, 1992:233–57.
5. Zelterman D, Louis TA. Contingency tables in medical studies. In: Bailar JC, Mosteller F, eds. Medical uses of statistics, 2nd ed. Boston, MA: NEJM Books, 1992:293–310.
6. Grodstein F, Stampfer MJ, Colditz GA, et al. Postmenopausal hormone therapy and mortality. New Engl J Med 1997;336:1769–75.
7. Fleiss JL. Statistical methods for rates and proportions, 2nd ed. New York: Wiley, 1981;sviii,321.
8. Grunkemeier GL, Payne N, Jin R, Handy JR Jr. Propensity score analysis of stroke after off-pump coronary artery bypass grafting. Ann Thorac Surg 2002;74:301–5.
9. McKinney WP, Young MJ, Hartz A, Lee MB. The inexact use of Fisher's exact test in six major medical journals. JAMA 1989;261:3430–3.
10. Moses LE, Emerson JD, Hosseini H. Analyzing data from ordered categories. In: Bailar JC, Mosteller F, eds. Medical uses of statistics, 2nd ed. Boston: NEJM Books, 1993:259–79.
11. Burns ER, Billett HH, Frater RW, Sisto DA. The preoperative bleeding time as a predictor of postoperative hemorrhage after cardiopulmonary bypass. J Thorac Cariovasc Surg 1986;92:310–2.
12. Blackstone EH. Comparing apples and oranges. J Thorac Cardiovasc Surg 2002;123:8–15.
13. Kitchell JR, Glover R, Kyle R. Bilateral internal mammary artery ligation for angina pectoris: preliminary clinical considerations. Am J Cardiol 1958;1:46–50.
14. Cobb L, Thomas G, Dillard D, et al. An evaluation of internal-mammary-artery ligation by a double-blind technique. N Engl J Med 1959;260:1115–8.
15. Klepetko W, Moritz A, Khunl-Brady G, et al. Implantation of the Duromedics bileaflet cardiac valve prosthesis in 400 patients. Ann Thorac Surg 1987;44:303–9.
16. Klepetko W, Moritz A, Mlczoch J, et al. Leaflet fracture in Edwards-Duromedics bileaflet valves. J Thorac Cardiovasc Surg 1989;97:90–4.