Interpreting Statistics in Medical Literature: A *Vade Mecum* for Surgeons

Ulrich Guller, MD, MHS, Elizabeth R DeLong, PhD

Background

For most of its history, the practice of medicine has been a profoundly empiric enterprise. Although this empiricism continues by necessity to exist in the clinical environment, the advent of scientifically rigorous epidemiology has transformed medical research in the 20th century.

The driving force behind the maturation of an epidemiologic approach to medicine has been the incorporation of statistical analysis in modern medical research, a practice that has become almost mandatory in past decades.[1,2] Sound statistical methods are essential to medical science because they transform ambiguous raw data into meaningful results.[3] But current trends toward evidence-based medicine can only flourish in a culture of statistical literacy. Such a culture requires physicians who are equipped with the necessary knowledge and skills to critically and accurately interpret statistical data.[2,4-6]

Unfortunately, there is ample evidence that many physicians are ill prepared to accurately interpret statistical computations in medical literature,[2,7,8] and, a significant association between number of years out of medical training and loss of statistical knowledge has been reported.[7] Given the ever-increasing prevalence of evidence-based practices, such a loss has potentially grave implications for the medical community.

This article provides a series of nontechnical explanations of basic statistical operations in medicine, coupled with intuitive examples drawn from the field of surgery. It is hoped that this *vade mecum* will facilitate the surgeon's critical appraisal of medical literature and its implementation in clinical practice.

No competing interests declared.

The Swiss National Foundation, Bern/Switzerland, Krebsliga beider Basel, Basel/Switzerland, Freiwillige Akademische Gesellschaft, Basel/Switzerland, and Fondazione Gustav e Ruth Jacob, Aranno/Switzerland provided financial support for Dr Guller's research fellowship at Duke University Medical Center. Partial funding for Dr DeLong was provided by a grant from the Agency for Healthcare Research and Quality (No. HS09940).

Received May 13, 2003; Revised September 17, 2003; Accepted September 18, 2003.

From the Departments of Surgery (Guller) and Biostatistics and Bioinformatics (DeLong), Duke University Medical Center, Durham, NC, and the Department of Surgery, Divisions of General Surgery and Surgical Research, University Hospital Basel, Basel, Switzerland (Guller).

Correspondence address: Ulrich Guller, MD, MHS, University of Basel, Department of Surgery, Divisions of General Surgery and Surgical Research, Spitalstr 21, CH-4031 Basel, Switzerland.

What is the difference between a mean and a median?

The center of a data distribution can be summarized by the mean or the median. The mean is the sum of all values divided by the number of observations. The median is the middle value when all observations are ranked from the least to the greatest (or vice versa). Why is it important to make a distinction between the mean and the median? The mean is sensitive to outliers (extreme values, data points that do not follow the pattern of most other data points); the median is not.

Example

Let us assume that we are evaluating patients after laparoscopic cholecystectomy. The primary end point (also known as the outcome or dependent variable) of the investigation is the length of hospital stay. For the sake of simplicity, let us say that our sample includes five patients. The lengths of hospital stay of the patients who all had a postoperative course without complications were 1 day, 1 day, 2 days, 3 days, and 3 days. In this example, the mean and median are identical (2 days). Now hypothesize that the fifth patient suffered from a postoperative infection that led to a generalized sepsis and respiratory failure requiring prolonged intubation, intravenous antibiotics, and transfer to the intensive care unit and that this patient was finally dismissed after 56 days of hospitalization. The mean length of hospital stay is now 12.6 days, but the median remains unchanged.

Outliers—patients who behave very differently from the majority of patients—are frequently present in medical literature and might render interpretation of the study findings difficult.[9] Use of the median helps prevent potential distortion of study findings caused by extreme values and should be preferably used if outliers are present.

ISSN 1072-7515/04/$30.00
doi:10.1016/j.jamcollsurg.2003.09.017

Abbreviations and Acronyms

CI = confidence interval
COPD = chronic obstructive pulmonary disease
IBD = inflammatory bowel disease
SD = standard deviation
SEM = standard error of the mean
VMA = vanillymandelic acid

But both mean and median sometimes fail to appropriately reflect the nature of the data. Consider a different sample of seven patients undergoing laparoscopic cholecystectomy. Let us assume that the lengths of hospital stays were 1 day, 1 day, 1 day, 2 days, 9 days, 11 days, and 14 days. The median in this example, 2 days, poorly summarizes the "middle" of the data. The mean, 5.6 days, doesn't give a clear picture of the nature of the data either. It is critical that a measure of the variability of the data and a mean or median be used to summarize a data distribution.

How to interpret standard deviations and standard errors of the mean

Although the standard deviation (SD) and standard error of the mean (SEM) are distinctly different statistics, they are often used interchangeably in medical literature. Confusion about their correct interpretation is considerable.[2,9]

The SD is a measure of the variability (scatter) of a data distribution. It is a measure of the degree to which the individual values deviate from the population mean. Larger deviations from the mean indicate more extensive scatter and result in larger standard deviations. The sample SD is an estimate of the population SD and is computed using deviations from the sample mean. There is a common misconception in the medical community about interpretation of the standard deviation: it is often believed that the standard deviation decreases with increasing sample size. Regardless of the sample size, the standard deviation will be large if the data are highly scattered.

The standard error of the mean (SEM) reflects the variability in the distribution of sample means from the population mean. If several different samples were available (which is not generally the case), their means would vary and would have a standard deviation, which is the SEM. Contrary to the SD, the SEM is an inferential statistic strongly dependent on the sample size.[9] The larger the sample size, the smaller the SEM, and the more precisely the sample mean estimates the overall population mean.

Because both sample SD and SEM are statistics frequently used in medical literature, it is important to know how to convert one to the other. The SEM can be obtained by dividing the SD by the square root of the number of patients in the sample ($SEM = SD/\sqrt{n}$). Multiplication of the SEM by the square root of the number of patients will result in the SD ($SD = SEM \times \sqrt{n}$). Again, the SD is the appropriate statistic to describe the scatter of the data; the SEM estimates the variability of an estimate of the sample mean. Some investigators display error bars using the sample-based estimate of the SEM instead of the SD for graphic presentation of the data scatter, leading readers to believe that there is little data dispersion.[2,9] The graphic presentation of SEM instead of SD to indicate data scatter is misleading.

A standard error can be computed not only for a mean but for any kind of sample statistic, eg, proportions, differences between means and proportions, regression parameters (as described below), and so forth. We will discuss the standard errors in context with confidence intervals.

How to interpret risk ratios, absolute risk reduction, odds ratios, and number needed to treat

Although the display and analysis of continuous outcomes (such as tumor size, length of hospital stay, total bilirubin concentration in the blood, and so forth) are easy and intuitive, the interpretation of percentages for dichotomous (binary, "yes or no") end points is often more challenging.[10] Examples of dichotomous outcomes are death, tumor relapse, liver failure, gastrointestinal bleeding, and so forth. Results of dichotomous end points are frequently presented in medical literature using risk ratios, odds ratios, absolute risk reduction, and the number needed to treat. Nonetheless, confusion exists in the medical community about the interpretation of these different analytical tools. To facilitate the understanding, let us consider the following hypothetical example.

There is extensive evidence in the medical literature that consumption of a low-fiber diet is a causative factor in the development of colon diverticulosis.[11-13] In the US, approximately one-third of people over 45 years of age and two-thirds of people over 85 years of age suffer

Table 1. Hypothetical Study of a Risk Factor and a Dichotomous Outcome

Risk factor (low-fiber diet)	Outcome (developing colon diverticulosis)				n
	Present		Absent		
Yes	A	90	10	B	100
No	C	40	60	D	100

from diverticular disease[13] leading to approximately 200,000 hospitalizations per year.[14] In Table 1, the risk factor is consumption of a low-fiber diet, and the outcome under investigation is the development of sigmoid diverticulosis. Let us assume that 40% of patients without the risk factor and 90% with the risk factor develop diverticulosis. The incidence of diverticulosis in patients with and without a low-fiber diet can be displayed in a 2 × 2 table.

Relative risk

The relative risk (also known as risk ratio, RR) is the likelihood of experiencing the outcomes in the group with the risk factor divided by the likelihood of experiencing the outcomes in the group without the risk factor. The relative risk [A/(A + B)]/[C(C/C + D)] in this example would be 0.9/0.4 = 2.25. That means that the risk of developing colon diverticulosis for patients with the risk factor is 2.25 times that of patients without the risk factor.

Absolute risk reduction

The absolute risk reduction equals the difference in the percentages of patients experiencing the outcomes in the patient subsets with and without the risk factor [A/(A + B) − [C/C + D)]. The absolute risk reduction represents the percentage of patients who did not have the adverse outcomes because of the absence of the risk factor. In our example, the absolute risk reduction is 0.9 − 0.4 = 0.5 or 50%. In other words, 50% of patients consuming a high-fiber diet do not develop sigmoid diverticulosis because they eat healthy, high-fiber diets.

The number needed to treat

The number needed to treat has been recently introduced into medical literature as a measure of treatment or prevention efficacy.[15] The number needed to treat (or in our example, the number needed to prevent) represents the number of patients that must be treated, or from whom a certain risk factor must be removed, to prevent the occurrence of one case. The number needed to treat is the inverse of the absolute risk reduction (in our example: 1/0.5 = 2). In other words, we would have to prevent two patients from eating low-fiber diets to prevent the development of sigmoid diverticulosis in one case.

Odds ratio

The odds are defined as the probability of experiencing an outcome divided by the probability of not experiencing the outcome.[16] All probabilities range from 0% to 100% but odds can be any positive number. The odds can be easily converted to probability, and vice versa:

Odds = probability of experiencing the outcome/
(1 − probability of experiencing the outcome)

Probability of experiencing the outcome = odds/(1 + odds)

An odds ratio (OR) can be computed by dividing the odds of patients exposed to the risk factor by the odds of patients without the risk factor. In Table 1, the odds ratio would be (A/B)/(C/D) or (0.9/0.1)/(0.4/0.6) = 13.5. Note that this calculation is equivalent to the ratio (A × D)/(B × C). The odds of developing sigmoid diverticulosis for patients with low-fiber diets are 13.5 times that of patients who have regular fiber intake. The odds ratio is the preferred method of displaying results for case-control studies, metaanalyses, and logistic regression analyses (discussion proceding).

Relative risk, absolute risk reduction, and odds ratio can be misleading because their clinical importance is highly dependent on the underlying prevalence of the disease. For instance, in our example, the relative risk of developing sigmoid diverticulosis in patients with low fiber intake is 2.25 times higher than that of patients with regular fiber consumption. Is this relative risk clinically relevant? That is dependent on the prevalence of the disease. In the US, where millions of people have sigmoid diverticulosis, the impact of regular fiber diet intake would have a large impact on the prevalence of this disease. Conversely, in some third world countries where the prevalence of diverticulosis is low, the same relative risk would be much less important.

Difference between the odds ratio and the relative risk

The RR is the intuitive measure of differential likelihood of disease. But some study designs preclude direct estimation of the RR. For example, suppose the diverticulitis study were performed as a case-control study in which 100 patients with diverticulitis and 100 disease-free controls were sampled for evidence of a low-fiber diet being a risk factor. The results might look like Table 2.

Table 2. Hypothetical Case-Control Study with a Dichotomous Outcome

	Outcome (developing colon diverticulosis)				
Risk factor (low-fiber diet)	**Present**		**Absent**		**n**
Yes	A	70	15	B	85
No	C	30	85	D	115

For these data, the odds ratio is $(A \times D)/(B \times C) = (70 \times 85)/(15 \times 30) = 13.2$, a similar figure to the one calculated from the previous table. But the apparent risk of disease for those without the risk factor is $30/115 = 0.26$, which is not an accurate representation of risk. Likewise, the apparent relative risk would be $(70/85)/(30/115) = 3.16$, which is an overestimate. The reason for this discrepancy is that the percentage of diverticulitis cases in this study is not representative of the prevalence of diverticulitis in the population. This study design allows an appropriate representation of the prevalence of the risk factor, but not the disease. But the odds ratio is not affected by the sampling design. For a relatively rare disease, the odds ratio is approximately equal to the relative risk.

How to interpret a confidence interval

The results of all studies are based on a limited number, or sample of patients. The findings of a study may or may not be representative of the overall population (target population) of patients chosen for the study. The goal of an investigator is to make statements that can be generalized from the study sample to all patients with the disease and the characteristics (eg, age, race, gender, and so forth) under investigation. To ensure that the findings in a study sample have a strict equivalence with the overall population parameters, a complete patient population, of very large or infinite size, would be required—an obvious logistic impossibility. In the absence of such methods, confidence intervals provide a useful tool to determine a range of values in which the parameters of the target population are likely to reside. A certain degree of confidence is chosen that indicates how sure the investigator is that the true value lies within the given range.

In the medical community, 95% confidence intervals (95% CI) are commonly used in the presentation of results. A 95% CI represents a range of values that will include the true population parameter in 95% of all cases. In other words, if you took an infinite number of samples of the same size, from the same overall population, and calculated the CI in the same manner, the true parameter in the overall population would be included 95% of the time. There remains a 5% chance that the true population parameter is outside the 95% CI. If an investigator wishes to be more sure that the confidence interval based on the patient sample includes the true population value, a 99% CI can be chosen, but the 99% CI is wider than a 95% CI because there is a smaller degree of uncertainty. The higher the level of confidence, the wider the confidence limits. A 95% CI can be computed for means, proportions, differences of means and proportions, risk ratios, odds ratios, sensitivity, specificity, and so forth. Again, computing confidence intervals only makes sense if the sample is representative of a larger population for which inferences can be drawn. The width of the confidence interval indicates the precision of an estimate and is dependent on the variation in the data and the number of subjects in the sample. The width of the 95% CI and the standard error (SE) are closely related: for samples of sufficiently large size ($n \geq 60$), the 95% CI is usually calculated as the mean $\pm 2 \times$ SEM. (The factor, with which the SEM is multiplied, varies with the samples size. For $n = 60$ it is exactly 2, for $n = 10$ it is 2.3, for an infinitely large sample size, the factor is 1.96. But the choice of a factor 2 is a good approximation in the vast majority of applications.) The greater the dispersion of data and the smaller the sample size, the larger the SE and the wider the confidence interval. Conversely, the less scattered the data and the larger the patient sample, the narrower the confidence interval. A wide confidence interval indicates that the sample data are insufficient for precisely estimating the effect in the overall population and must be interpreted cautiously, regardless of whether or not the results are statistically significant.[17,18]

Example

Twenty-four hour urine measurement of vanillylmandelic acid (VMA) represents a sensitive and specific test in the diagnosis of pheochromocytoma patients.[19,20] Let us consider a hypothetical sample of 10 patients with pheochromocytoma. The urine measurements in these patients yielded VMA values of 50, 60, 70, 80, 90, 100, 110, 120, 130, and 140 mg/24 h (normal: <7 mg/24 h). The 95% CI for the mean VMA value ranges from 73.3 to 116.7 mg/24 h (sometimes displayed as: 95% CI (73.3, 116.7 mg/24 h). The first value is called the lower

and the second value the upper confidence limit. We can be 95% confident that the interval 73.3 to 116.7 mg/24 h contains the true population mean for pheochromocytoma patients. If our sample size were five times larger and had the same range and value distribution, the 95% CI would be (86.8, 103.2 mg/24 h). A sample size 10 times larger would result in a 95% CI of (89.2, 100.8 mg/24 h). Again, the larger the sample size, the narrower the 95% CI, and the more precise the sample estimate. If we had a sample of 10 patients with less variability than in our previous sample (mean VMA values: 91, 92, 93, 94, 95, 95, 96, 97, 98, 99), the 95% CI would be much smaller (93.1, 96.9). This simple example shows that the main determinants of the width of a confidence interval are the sample size and the data dispersion.

How to interpret type I and type II errors, sample size computations, and power

The findings of a study comparing two groups of patients can be wrong in two ways[21]:

1. The results might lead to the erroneous conclusion that there is a difference between the study groups when, in reality, there is none.
2. The results might lead to the erroneous conclusion that there is no difference between the study groups when, in reality, a difference exists.

The first situation represents a false-positive result and is called a type I error. The bound that we put on the probability of committing a type I error is named *alpha*.[21] *Alpha* is also referred to as the level of statistical significance or significance level. Number 2 in the preceding list represents a false-negative result and is called a type II error. The probability of committing a type II error is referred to as *beta*.[21,22]

An *alpha* of 0.05 is commonly used in medical research. This means that a 5% chance of obtaining a false-positive result is considered acceptable. *Alpha* is the benchmark to which p values (discussed later in the article) are compared. If the p value is larger than *alpha,* a result is said to be nonsignificant. On the other hand, if the p value is smaller than the benchmark *alpha,* the findings are statistically significant. In other words, *alpha* is the threshold p value below which a result is called statistically significant. *Beta,* the false-negative rate, is complementary to the power of a study. In medical science, *beta* is commonly assumed to be at a level of 0.2 or 0.1, indicating a power of 80% or 90%, respectively. Power is defined as the probability of finding a statistically significant result (of rejecting the null hypothesis) in a study, if the populations are truly different.[21] The choice of adequate power in a study is critical because investigators and funding agencies must be confident that an existing difference in the overall population can be detected using the study sample. If, for instance, the power in a randomized controlled trial is set at 90% (*beta* of 10%) and a true difference exists between the study arms, we would be able to detect that difference in 9 out of 10 cases if the trial were repeated an infinite number of times.

The power of a study is dependent on the following factors[21,22]:

1. The extent of the true difference between the populations under investigation
2. The *alpha* level (accepted rate of false-positive results)
3. The sample size

With larger sample sizes, larger true differences between the populations from which the patient samples have been drawn, or higher acceptance of false-positive results, the power of the study increases.

Before initiating the study, power and sample size must be determined. For sample size computations, investigators start by defining a clinically meaningful difference between treatments A and B, which is believed to be true for the overall patient population. This difference is usually based on preliminary data of small phase II studies or retrospective reviews, but is sometimes specified according to clinical intuition. If the investigator is satisfied with an 80% probability of obtaining a statistically significant difference between the study groups, if such a difference truly exists, a smaller sample size is required than if 90% power were chosen. In other words, a larger sample size corresponds to a higher level of power. Ideally, both *alpha* and *beta* would be set at 0 to avoid false-positive and false-negative findings. This would require a prohibitively large sample size, rendering any trial unfeasible. For a patient sample of given size, there is a tradeoff between *alpha* and *beta:* the more stringent *alpha* (the lower the false-positive rate), the higher *beta* (increased rate of false-negative results, lower power), and vice versa.[21,23] In general, one should choose a small *alpha* level if avoiding false-positive results is particularly important (eg, testing the efficacy of a new chemotherapy regimen with serious adverse effects). Similarly, a small *beta* level should be chosen if obtaining a false-negative result would be deleterious; for example,

Table 3. Sample Size Computations* for a Study Showing the 5-Year Overall Survival Difference Between Colorectal Cancer Patients With and Without Disseminated Tumor Cells

Expected 5-y overall survival of patients with disseminated tumor cells (%)	Expected 5-y overall survival of patients without disseminated tumor cells (%)	*Alpha*	*Beta*	Total sample size
45	75	0.05	0.20	79
45	75	0.05	0.15	89
45	75	0.05	0.10	105
50	75	0.05	0.20	108
50	75	0.05	0.15	124
50	75	0.05	0.10	145
55	75	0.05	0.20	161
55	75	0.05	0.15	185
55	75	0.05	0.10	215
55	70	0.05	0.20	287
55	70	0.05	0.15	328
55	70	0.05	0.10	384

*All sample size computations are based on a 25% positivity rate of the samples, a type I error probability of 0.05 (two-sided), a projected accrual period of 3 y, and a followup interval of 5 y. The program used for sample size computations was based on references 61 through 65.

if an investigator wants to demonstrate the superiority of a new, less invasive surgical procedure over an established procedure associated with considerable short- and longterm sequelae. The false conclusion that the new procedure is not as effective as the standard procedure would put new patients at risk of suffering worse outcomes.

It is imperative that the authors of a clinical trial report the parameters on which the computed sample size is based.[24,25] Despite this, many investigators fail to do so.[22,25,26] If no information about power calculations is reported, the reader does not know if:

1. No sample size requirement was computed.
2. The investigators were unable to accrue the initially computed patient number.
3. The trial was extended beyond the initially computed sample size to obtain higher statistical power.
4. The investigators stopped the trial earlier than anticipated because the interim results were favorable.[26]

Example: table 3

There is suggestive evidence in the medical literature that colorectal cancer patients with single disseminated tumor cells in bone marrow and peritoneal lavage samples have a higher risk of suffering a relapse and a shorter overall survival compared with patients without disseminated tumor cells.[27-29] Let us say that we want to design a study that allows us to evaluate the prognostic significance of disseminated tumor cells in colorectal cancer patients. Assuming the 5-year overall survivals for stages I and II colorectal cancer patients with and without disseminated tumor cells to be 45% and 75%, an *alpha* level of 0.05, and a power of 80%, the required number of patients is 79 (Table 3). Table 3 displays sample size computations that would answer the same research question using different overall survival and power estimates. It is important to realize that sample sizes are highly dependent on the assumed estimated difference in survival rate and the chosen power.

How to interpret a p value

Remember that statistics help us to make inferences from the patient sample under investigation to the overall population. To understand the meaning of a p value, it is necessary to understand the meanings of null and alternative hypotheses. The null hypothesis of a study often is the hypothesis that no difference exists between the study groups. In a randomized clinical trial, for example, the null hypothesis states that there is no difference between study arms for the end point under investigation (eg, disease-free or overall survival, postoperative complications, postoperative mortality, and so forth). Conversely, the alternative hypothesis (the one the investigator wants to demonstrate) is that there is a significant difference between study arms. Let us assume that in the overall population, the end points for patients assigned to arms 1 and 2 of a two-armed randomized clinical trial are identical and the intervention has no effect. Nonetheless, it is possible that certain patients respond more favorably to the intervention than others. As we deal with a sample of the overall population it can

be hypothesized that a difference might occur because of chance alone (eg, sampling variations). The p value is the probability that the difference between arms 1 and 2 is at least as large as that observed in the sample if there is actually no difference in the overall population (assuming the null hypothesis).

Example

Let us consider a randomized clinical trial comparing preoperative radiation therapy plus surgery (arm 1) versus surgery alone (arm 2) in the treatment of resectable esophageal cancer. Assuming that overall survival is the primary end point, the null hypothesis of the investigation states that survival time in both arms is the same. Conversely, the alternative hypothesis claims a survival difference between patients randomized to arm 1 and those randomized to arm 2. Let us suppose that after completing intervention and followup, patients in arms 1 and 2 had median overall survivals of 18 months and 22 months, respectively, and that the p value for this survival difference was $p = 0.02$. Interpretation of this result is: If the new intervention has equivalent overall survival to the standard procedure (if the null hypothesis were true), there is a 2% chance of observing a survival difference as large as or larger than the one observed. In other words, if truly there were no difference between the treatments in the overall population, and the trial were repeated an infinite number of times, an overall survival difference of 4 months or more would be expected to occur by chance in only 2 of every 100 such trials. If the p value is small, the probability of obtaining the observed difference by chance alone is low, and one usually assumes that the null hypothesis does not hold. Conversely, if the p value is large, it is conceivable that the data are consistent with the null hypothesis, which cannot be rejected.

The following issues about the interpretation of a p value are of prime importance:

1. A p value is the probability of getting a difference at least as large as the one observed, under the assumption that the null hypothesis is true (assuming that there is no difference between the populations under investigation). A p value without a null hypothesis is meaningless. One should never interpret a p value without knowing the null hypothesis with which it is associated.
2. A highly significant p value (eg, $p = 0.001$) tells you that the difference observed in your study would occur very rarely (in only 0.1% of all cases) if truly there were no difference between the study groups. The p value does not prove that the alternative hypothesis is true. p Values are based on the assumption that the null hypothesis is true and only provide evidence against the null hypothesis, not evidence to support the alternative hypothesis.
3. The p value depends on the existing difference between the study groups, the scatter of the data (the standard deviation), and the sample size. The larger the difference between the study groups, the smaller the standard deviation, or the larger the sample size, the more significant the p value. In light of these factors that influence the magnitude of the p value, the benchmark of 0.05 should not be used as a clear cutoff between relevant and unimportant results. Guyatt and colleagues[30] emphasized this fallacy: "Why use a single cut-off point [for statistical significance] when the choice of such a point is arbitrary? Why make the question of whether a treatment is effective a dichotomy (a yes-no decision) when it would be more appropriate to view it as a continuum?"
4. A nonsignificant p value does not demonstrate that the null hypothesis is true. As mentioned previously, large p values might be simply due to small sample sizes or highly scattered data. A nonsignificant p value tells you only that the evidence is not strong enough to reject the null hypothesis.[31,32]
5. A p value is claimed to be statistically significant if it is smaller than the threshold of statistical significance (*alpha*). The latter is most commonly set at 0.05, but, in certain situations can be lower (see multiple comparisons).
6. Frequently researchers make statements such as "the association was found to be statistically significant (p value < 0.05)." What does this mean? The p value could be 0.049 or 0.00001. It is much more informative and helpful to the reader to give the exact p value and even better to display the confidence interval.[17]

Despite the fact that a p value of 0.05 is frequently considered a default benchmark for significant results, the instances discussed earlier show that this is a fallacious standard. Interpreting a p value is sensitive to a host of factors, all of which should be taken into account by a conscientious researcher.

How to interpret one-tailed versus two-tailed p values

An investigator who compares a new treatment to the standard treatment may have reason to believe that the new therapy is superior based on phase II studies or retrospective reviews. Should a one-tailed (one-sided) or two-tailed (two-sided) p value be used to compare these treatments? Both one- and two-tailed p values are based on the null hypothesis (that the treatments are equally

effective). A two-tailed (or two-sided) p value represents the probability that the difference between two treatments—assuming the null hypothesis to be true—is as large or larger than observed, with either treatment being superior to the other. Conversely, a one-tailed (one-sided) p value represents the probability that the difference observed would have occurred by chance alone, with one treatment being superior to the other as specified in the alternative hypothesis.[33] The one-tailed p value is usually half of the two-tailed p value. Although two-tailed p values are commonly used throughout medical literature, some investigators argue that one-tailed p values are appropriate in certain situations. Here are some general guidelines about this issue:

1. Unless you can state with absolute certainty that a difference between two interventions can only go in one direction, a two-tailed p value should be used.[33] For instance, although you might believe that a new radiochemotherapy regimen for rectal cancer patients does improve overall survival, patients might actually die earlier from unexpectedly severe side effects.
2. If you use a one-tailed p value, the alternative hypothesis must be stated in advance (a priori hypothesis), specifying the intervention believed to be superior.[34]
3. There have been instances in medical science when, at the end of a trial yielding a marginally significant two-tailed p value (eg, 0.06) for the difference between interventions, the p value was switched to a one-tailed p value (0.03) to obtain a statistically significant result. Such behavior is misleading and should be abandoned. Some authors have suggested that the level of statistical significance should be set at 0.025 if a one-tailed p value is used.[35,36]

Clinical versus statistical significance of a result

As previously mentioned, the magnitude of the p value depends on the sample size. If the sample size is large, even tiny differences between study groups will become statistically significant. The question is whether these small differences are clinically relevant. Statistically significant results may well prove to be trivial.[3] On the other hand, even though the p value might not be statistically significant (eg, from a small sample size), the differences found between the study groups might appear to be clinically relevant.[3,22] Frequently, only p values are reported in the medical literature but they lose their relevance if the sample sizes are large. In these situations, confidence intervals are helpful and informative in interpreting study findings and should be provided in addition to p values.

Example

Let us compare the mean length of hospital stay after open versus laparoscopic appendectomy. We will hypothesize that the true mean length of hospital stay is 3.16 days for patients undergoing laparoscopic surgery and 3.20 days for patients having open appendectomy and that the standard deviation for length of stay is 0.5 days. Is this difference of any clinical relevance? Almost certainly not, as the difference (0.04 days) is only approximately 1 hour. If, 4,908 patients or more (Sample size computation based on *alpha* of 0.05, *beta* of 0.2, and standard deviation of 0.5 days) are evaluated, half of them undergoing open appendectomy and the other half having laparoscopic surgery, the p value will become significant at a level of 0.05. If 10,678 patients (and this large patient number is not uncommon for retrospective secondary data analyses) are in the study, the p value becomes highly significant (0.001), and with 14,004 patients the p value becomes extremely significant (0.0001). Conversely, let's assume that laparoscopic appendectomy is truly associated with a shorter length of hospital stay compared with the open procedure and that the difference is half a day (3.2 days for laparoscopic appendectomy and 3.7 days for open appendectomy). Considering that length of hospital stay is correlated to hospital costs,[37,38] a difference in length of stay of 0.5 days between patients undergoing open and laparoscopic appendectomy undoubtedly represents a clinically important finding. If less than 34 patients are evaluated, the p value will not be statistically significant at an *alpha* level of 0.05, demonstrating once again that one of the main determinants of statistical significance is sample size.

How to view confidence intervals and p values as being complementary

Researchers are more likely to report p values than confidence intervals.[26,30] But confidence intervals provide much more information to the astute reader than do p values alone[3,17,26] and are now requested by many journals in the reporting of study findings.[1,39] Although p values and confidence intervals might seem different at first glance, closer scrutiny reveals that they are complementary. Both are computed using the same underlying assumptions. If the 95% confidence interval includes the value of the null hypothesis (eg, 0 for the difference between means, or 1 for a risk ratio and or odds ratio), the p value will be greater than 0.05. On the other hand,

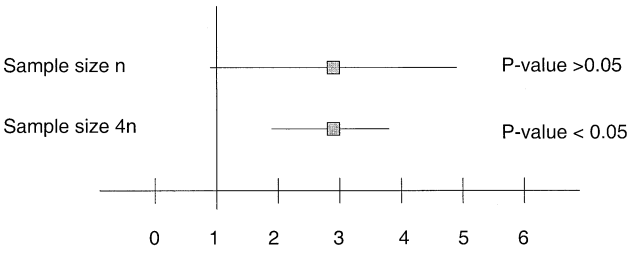

Figure 1. Point estimates and 95% confidence intervals for two samples. If the 95% confidence interval includes 1, the p value is not significant.

if the 95% CI does not include the value of no difference, the p value will be less than 0.05. Confidence intervals are also very helpful in interpreting nonsignificant p values.[3,17] If the range of the 95% CI of differences or risk ratios includes values that are clinically trivial, one can assume the results to be irrelevant with higher confidence. If the 95% CI includes values that you find clinically important, the study should be considered inconclusive because a small sample size might be a reason for not reaching statistical significance.

Example

It is well known that smokers are at higher risk for gastric ulcers compared with nonsmokers.[40-42] Let us assume that a study comparing the incidence of gastric ulcers of smokers versus nonsmokers found a relative risk and a 95% confidence interval of 2.9 (0.9, 4.9). Because the 95% confidence interval includes 1 (the value of the null hypothesis), the corresponding p value will be nonsignificant. As explained previously, this does not necessarily demonstrate that smokers do not have an increased risk for gastric ulcers compared with non-smokers. The width of the 95% confidence interval is considerable (because of a small sample size), ranging from a relative risk of 0.9 to 4.9. So smokers might be 0.9 times through 4.9 times more likely to develop gastric ulcers compared with nonsmokers. This confidence interval certainly covers values that have clinical importance, so, the findings of this hypothetical study should not be declared negative but rather inconclusive. Let's say that we had assessed a sample size four times larger than the initial study population, the 95% confidence interval would have narrowed to (1.9, 3.9), and, because the null value is not included anymore, the p value would have been significant. This example is graphically displayed in Figure 1.

Beware of multiple comparisons and subset analyses

If no difference exists between two groups of patients in the overall population (if the null hypothesis were true), the p value tells you how likely it is to get a difference at least as large as observed by coincidence. As explained previously, to limit the probability of a false-positive (type I) error, the threshold of statistical significance (*alpha* level) is usually set at 0.05. If you test multiple independent null hypotheses—all of which are true—in the same investigation, the probability that one p value might become statistically significant by chance alone increases. Multiple comparisons carry the risk of providing false-positive results.[32,43]

Most investigations in the medical literature test many different null hypotheses,[26,44-46] so the probability of getting a statistically significant result by coincidence is likely to exceed the standard of 5%. For instance, if you test 10 different independent null hypotheses, the probability of obtaining a statistically significant result at an *alpha* level of 0.05 is 40% ($1–0.95^{10}$), for 50 null hypotheses it is 92% ($1–0.95^{50}$), for 100 null hypotheses over 99% ($1–0.95^{100}$).[43] (The numbers in parentheses represent the formulae to compute the risk of obtaining a false-positive result.) p Values must be interpreted cautiously if many independent null hypotheses are tested.

Often measures are taken to decrease the risk of obtaining false-positive results caused by multiple comparisons. For instance, the Bonferroni method[23]—the simplest and most often used technique to adjust for multiple comparisons—divides the *alpha* level by the number of independent hypotheses tested.[47] If you test 5 hypotheses, the level of statistical significance should be decreased to 0.01 (0.05/5); if you test 10 different null hypotheses, the level of statistical significance should become 0.005 (0.05/10). The Bonferroni method has conservative properties and should not be used for adjusting for more than 10 hypotheses.[44]

Subset analyses are common in medical literature and are similar to multiple comparisons about their potential risk of obtaining spurious results, and their requirement for cautions interpretation.[45,48] After evaluation of the outcomes in the overall sample, study findings are assessed in subsets of patients (eg, stratified for age, gender, preexisting risk factors, severity of the disease, and so forth). Many investigators perform subset analyses regardless of the overall outcomes of the study. If the overall outcomes of a trial show a significant difference be-

tween study groups, subset analyses might be performed to identify patients who particularly benefit from the treatment. Conversely, if the overall outcomes in a study are negative (no statistically significant difference between the study groups), subset analyses are frequently performed to show some benefit of the treatment in at least a certain subset of the patients. The dangers of performing subset analyses are well known.[32] As discussed previously, the increased rate of false positivity derives from making multiple comparisons, testing multiple different hypotheses, and as a result, getting statistically significant results at a level that exceeds *alpha* even if all tested null hypotheses were true. If, on the other hand, an investigator finds nonsignificant results in subgroups, the conclusion that in reality there is no difference within these subsets might be erroneous too, because the study was not designed and adequately powered to detect a significant difference in the patient subsets. So, subgroup findings should be viewed as exploratory, subject to confirmation in another trial.

There is ongoing debate in the medical literature about whether, and how, to adjust for multiple comparisons.[23,44,47,49] The following guidelines are crucial in the interpretation of study subgroup analyses:

1. A priori versus a posteriori hypotheses: It is important to make the distinction between hypotheses that were created before performing the study (a priori) versus hypotheses that were stated after the conduct of the study (a posteriori).[26,47] A priori stated hypotheses do not carry the risk that the investigator was influenced by the readily available data, so are less prone to erroneous conclusions. If hypotheses are stated a posteriori, it is possible that the investigator looked at different patient subsets until significant results were found. This phenomenon is often referred to as "data mining," "data dredging," or "fishing expedition." Investigations for which hypotheses are formulated after the study has been conducted should be viewed more as hypothesis generating rather than hypothesis testing, and even more so if they look at patient subsets and perform multiple comparisons.[44,47,50]
2. Often investigators test multiple hypotheses, but report only the statistically significant findings.[47] This is misleading and potentially dangerous if the study has clinical implications. It is imperative to clearly state and report all tested hypotheses and associated p values.

Example

Inflammatory bowel disease (IBD) represents a public health problem affecting approximately 1 of 1,000 individuals in Western countries. The etiology and pathogenesis of IBD await additional clarification. It is generally accepted that IBD occurs predominantly in genetically susceptible people[51,52] yet many genetic loci involved in the pathogenesis of this disease need to be determined.[53] Satsangi and colleagues[54] performed a genome-wide search for susceptibility genes in patients with IBD. The authors investigated 260 different genetic loci for potential linkage to IBD. Because 260 null hypotheses were tested, there was a definite risk of obtaining false-positive results if the standard *alpha* level of 0.05 was used. So the investigators correctly chose a level of statistical significance of $p < 0.001$, which is 50 times smaller than the commonly used cutoff ($p < 0.05$). Also, the investigators viewed their study as hypothesis generating rather than hypothesis testing.

How to interpret survival curves

A survival curve is a graphic presentation of time to event data. The term *survival curve* is somewhat misleading, as not only time to death, but time to any event can be graphically displayed.[55] For instance, a "survival" curve can plot time to tumor recurrence, time to extubation after a surgical procedure, time to rejection of a kidney transplant, and so forth. The starting point of a survival curve, or time zero, marks the beginning of the period of observation for the event under investigation. For instance, time zero can be when the patient enters a protocol, the time point of randomization, the day of operation, start of adjuvant chemotherapy, and so forth. By definition, the start of a survival curve is always set at 100% because all patients whose clinical course is graphically displayed are alive at this point in time. Each step down on a survival curve represents the occurrence of an event. If two (three) patients experience the event under investigation, the step down is twice (three times) as large as that for one event.

In the majority of survival analyses, some patients are "lost" before the event occurs, or before followup is complete for the study. This phenomenon occurs frequently in investigations that enroll patients over many years. So the length of followup varies greatly among patients. Patients who do not experience the event of interest are referred to as censored, either during the study (eg, because of loss to followup) or at the end of the study.[56] Censoring enables patients to provide valuable information to the study despite not being followed over the entire investigation. When a patient is censored, the sur-

vival curve remains horizontal. The number of patients at risk decreases, so, the next event will result in a larger step down than the previous one.

It is important that survival curves indicate when a subject is being censored. Graphically this is usually done using tick marks. Display of censored subjects enables the reader to deduce how the number of patients at risk has decreased.

To compute the proportion of patients surviving through a given day, one needs to divide the number of patients at risk at the end of a day by the number of patients at risk at the beginning of the day, excluding censored patients both from the numerator and the denominator. To compute the proportion of patients surviving over a certain time period, multiply the proportion of patients surviving day 1, day 2, day 3, and so forth.[36]

As with means, proportions, and differences in means and proportions, one can compute the 95% confidence intervals around survival curves.[55] Although any survival curve is a graphic presentation of time to event data of a finite sample, as with any statistic, the true population survival curve likely differs from that of the sample. We can be 95% sure that the true survival curve of the overall population lies within the 95% confidence interval limits. The smaller the sample size, the larger the confidence limits around the survival curve.

The survival curve allows us to easily obtain the median survival (or time to recurrence, and so forth) by crossing a horizontal line through the survival curve at the 50% mark of the ordinate. The number of months or years at the abscissa of this crossing point represents the median survival time (Fig. 2). If the horizontal line drawn from the 50% mark of the ordinate does not cross the survival curve, less than half the patients have experienced the event and the median time to event cannot yet be computed.

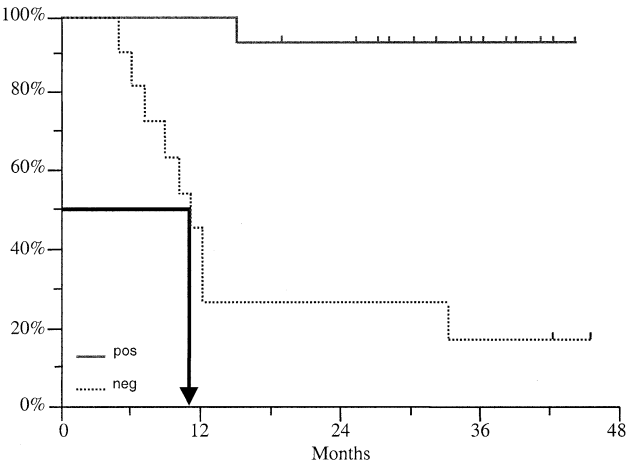

Figure 2. Hypothetical survival curve comparing two groups of colorectal cancer patients. The solid line survival curve plots time to death for node negative, the dotted curve for node-positive patients. In the group without lymph node metastases, only one patient dies at 15 mo of followup. Nine node-negative patients die, two (step is twice as big) at 12 mo of followup. Median survival for the node-positive group is approximately 10 mo; the median survival for the negative group cannot be computed because fewer than half of the subjects are deceased at the end of the study. Note that 16 patients were censored, and that the survival curve remains horizontal during censoring.

What is a confounding variable?

Before defining a confounding variable it is important to understand the meaning of, and the association between, a predictor variable and an outcome. Commonly, studies are designed to show a link between a predictor variable (independent variable) and an outcome (dependent variable). Predictor variables can be either a diagnostic or therapeutic interventions (eg, new surgical therapy, new diagnostic procedure) or a risk or prognostic factor such as age, patient comorbidity, tumor size, lymph node status, and so forth. Frequently assessed outcomes in surgical literature are disease-free survival, overall survival, response to a treatment, and postoperative morbidity. A confounding variable (confounding factor, confounder) is an extrinsic factor that is linked to the predictor variable and also impacts the outcome. The perceived association between the predictor and the outcome variable is distorted because of the confounder.[57]

Example

An intuitive example to illustrate confounding is the relationship of frequent bar visits to the development of liver cirrhosis. We all know that frequent bar visits in and of themselves are of no danger to the liver if the person going to the bar only consumes soft drinks and watches a basketball game. It is obvious that frequent bar visitors have a tendency to consume alcohol, which is unequivocally linked to developing liver cirrhosis.[58,59] Alcohol consumption in this case is the confounding variable (Fig. 3). If we categorize a sample of people by whether they are frequent or nonfrequent bar visitors and assess the relationship of this variable with liver cirrhosis, we most likely will find a high degree of association. But if we stratify the population of frequent bar visitors by the confounding factor (alcohol consumption), frequent bar

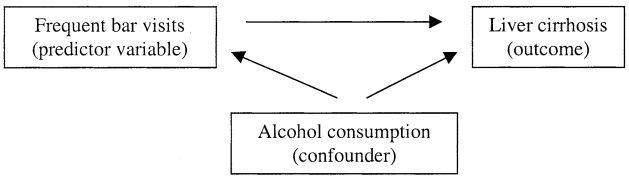

Figure 3. Relationship between putative risk factor, confounding variable, and outcome.

visits would no longer be associated with developing liver cirrhosis.

How to interpret multivariable analyses

In medical literature—especially in nonrandomized observational studies—confounding can be a major problem. For instance, in an observational study comparing two treatments, patients undergoing treatment A might substantially differ from patients having treatment B with respect to factors such as age, gender, race, socioeconomic status, comorbid diseases, and tumor staging and grading. All of these are probably also related to the outcomes of interest, so are potential confounding variables. The investigator seeks to assess the relationship between the primary predictor variable (type of therapy, A versus B) and the outcomes under investigation after the potential distortion through covariates has been eliminated. The use of stratification when multiple potential confounders are present is cumbersome. The preferred method of adjusting for many confounding variables simultaneously is multivariable analyses.

Depending on the type of outcome variable, one of three different multivariable analyses is generally appropriate: For continuous outcomes (eg, postoperative length of hospital stay), multiple linear regression analysis is appropriate; for categorical or dichotomous outcomes (eg, presence of metastases), multiple logistic regression analysis is useful; and for time to event outcomes (eg, time to death or time to recurrence), the proportional hazards regression analysis is often chosen (Table 4). For any outcome, multivariable analyses provide the risk-adjusted (often called "independent") impact of the primary predictor variable on the outcomes after controlling (risk-adjusting) for the potential confounding of all other covariates.

Example

Consider a prospective observational study comparing open versus laparoscopic colectomy for sigmoid diverticulitis with end points such as postoperative morbidity and mortality, length of hospital stay, operating time, and cost. Let us assume that study findings show laparoscopic colectomy to be clearly superior to open colectomy for these end points. Does this mean that laparoscopic colectomy is truly better than open surgery? Not necessarily. It is conceivable that patients undergoing open versus laparoscopic colectomy differ in important risk factors. Patients receiving open colectomy might be older, sicker, have more complicated disease, have lower socioeconomic status, or be operated on by less experienced surgeons. Any of these factors might confound the relationship between the primary predictor variable (type of procedure: open versus laparoscopic colectomy) and the end points. To obtain the true benefit (if any) of laparoscopic colectomy, multivariable analyses must be performed. All potential confounding variables must be included in the statistical models to adjust for the differences between the patient subsets undergoing open and laparoscopic procedures. Again, use of multivariable analysis is particularly important in nonrandomized trials because imbalances between the study groups can be expected. Randomization helps to distribute known and unknown confounding variables equally among arms, so analyses of a randomized controlled trial do not necessarily require multivariable adjustment.

Table 4. How to Interpret Positive and Negative Beta Coefficients*

Type of multivariable analysis	Positive beta coefficient	Negative beta coefficient
Multiple linear regression analysis	Mean value of the outcome increases with presence of risk factor (if dichotomous) or as independent continuous variable increases.	Mean value of the outcome decreases with presence of risk factor (if dichotomous) or as independent continuous variable increases.
Multiple logistic regression analysis	Probability of the outcome increases with presence of risk factor (if dichotomous) or as independent continuous variable increases.	Probability of the outcome decreases with presence of risk factor (if dichotomous) or as independent continuous variable increases.
Proportional hazard regression analysis	Hazard increases with presence of risk factor (if dichotomous) or as independent continuous variable increases.	Hazard decreases with presence of risk factor (if dichotomous) or as independent continuous variable increases.

*Presence of risk factors is coded as 1, lack of risk factor as 0.

Table 5. Hypothetical Multiple Linear Regression Model: Comparison of Length of Hospital Stay Between Open and Laparoscopic Appendectomy

Variable	Beta coefficient	Standard error	p Value	95% Confidence interval
Laparoscopic surgery	−0.67	0.04	<0.0001	[−0.74, −0.61]
Age (y)	0.02	0.001	<0.0001	[0.019, 0.024]
COPD	0.37	0.04	<0.0001	[0.29, 0.45]

COPD, chronic obstructive pulmonary disease.

In all three types of multivariable analysis, the variable's beta coefficient (not to be confused with the previously mentioned rate of false-negative results) indicates how the dependent variable responds to changes of the independent variable, after adjusting for all other covariates in the model. As the type of model changes for different types of end points, the interpretation of the beta coefficient changes too (Table 4).

In multiple linear regression analysis, the outcome is continuous. A positive beta coefficient signifies that the independent variable and the mean value of the dependent variable vary in the same direction (either both increasing or both decreasing). Conversely, a negative coefficient indicates that while the independent variable increases the mean outcome decreases, and vice versa.

In multiple logistic regression analysis, the appropriate statistical tool to assess the risk-adjusted impact of covariates on categoric end points, the logit is modeled. The logit is the natural logarithm of the odds of experiencing the outcomes. The beta coefficient in logistic regression models expresses how the logit changes with changes in the independent variable. Although the logit is difficult to interpret on its own, it can be easily transformed into the odds ratio by taking the antilogarithm (exponentiating) of the beta coefficient.

As mentioned previously, proportional hazards regression is frequently used if a time to event end point is being evaluated. Time to event outcomes incorporate more information than dichotomous end points if the followup period is sufficiently long (eg, we not only know whether or not death occurred but also how long after cancer resection).[32,36,60] In proportional hazards regression analysis, the beta coefficient represents the change in the natural logarithm of the relative hazard for a one-unit change in the independent variable. The relative hazard represents the ratio of instantaneous risk of experiencing the event for patients having a certain risk factor to the instantaneous risk of experiencing the event for patients where this risk factor is absent. Similar to the coefficients in logistic regression analyses, the relative hazard is obtained by taking the antilogarithm of the beta coefficient.

It is important to remember that larger beta coefficients, or smaller standard errors result in more significant p values. This applies to all three multivariable techniques. As a rule of thumb, a beta coefficient that exceeds its standard error by a factor of two will likely be associated with a statistically significant result. It seems intuitive that an estimate associated with a large standard error (which, as discussed previously, represents the precision of the estimate) is unlikely to be statistically significant.

Example 1 (table 5)

Let us assume that we are performing a retrospective analysis of open and laparoscopic appendectomy. The outcome under investigation is length of hospital stay measured in days, and the main predictor variable is the type of procedure (laparoscopic versus open appendectomy). Because we are dealing with a retrospective review rather than a prospective randomized clinical trial, it is conceivable that patients undergoing laparoscopic operation were substantially different from open appendectomy patients in regard to putative confounders such as age and presence of chronic obstructive pulmonary disease (COPD). To obtain the risk-adjusted difference in length of hospital stay between open and laparoscopic appendectomy, we must include age and presence of COPD in the model.

Laparoscopic surgery. Let us remember that multiple linear regression models the mean value of the outcomes. The beta coefficient for laparoscopic surgery is −0.67, meaning that the mean length of hospital stay of patients undergoing laparoscopic appendectomy is 0.67 days shorter (because of the negative sign) than that for patients undergoing open procedures after adjusting for potential age and race differences. In this example, laparoscopic appendectomy was coded as 1 and open appendectomy as 0. The importance of correct interpretation of coding becomes evident in this example: If open ap-

Table 6. Hypothetical Multiple Logistic Regression Model: Comparison of Postoperative Infections Between Open and Laparoscopic Appendectomy

Variable	Beta coefficient	Standard error	p Value	Odds ratio and 95% confidence interval
Laparoscopic surgery	−0.535	0.05	<0.0001	0.59 [0.50, 0.69]
Age (y)	0.007	0.001	<0.0001	1.007 [1.004, 1.01]
COPD	0.10	0.06	0.1	1.11 [0.96, 1.32]

COPD, chronic obstructive pulmonary disease.

pendectomy were coded as 1 and laparoscopic appendectomy as 0, the beta coefficient would have been +0.67, meaning that open appendectomy patients have a length of hospital stay 0.67 days longer than laparoscopic appendectomy patients.

Age. The beta coefficient for age is 0.02. So, for each year of age, the average length of hospital stay increases by 0.02 days. In other words, 60-year-old patients are expected to have a 0.8-day (40 × 0.02 days/year = 0.8 days) longer length of hospital stay than 20-year-old patients.

Chronic obstructive pulmonary disease (COPD). COPD was coded as present (1) and absent (0). The presence of COPD is associated with hospital stays 0.37 days longer than those for patients who do not have COPD. Again, the interpretation of coding is critical. If absence of COPD were coded as 1, the beta coefficient would have been −0.37.

For all three covariates in this model, the beta coefficients are more than twice as large as the standard errors, so the p values are statistically significant. Also, the 95% confidence intervals do not include the value of the null hypothesis (for continuous outcomes = 0), indicating that the p values are below the level of statistical significance.

Example (table 6)

Let us consider the same study evaluating a different outcome (postoperative infections). As a postoperative infection either occurs or doesn't (dichotomous outcome), the multiple logistic regression model is the appropriate statistical tool for the analysis. Remember that multiple logistic regression analyses model the logit of the outcomes (natural logarithm of the odds of experiencing the outcomes). The logit is difficult to interpret. Thankfully, the logit can be easily transformed into the odds ratio by taking the antilogarithm (exponentiating). If the beta coefficient in a logistic regression model (and in a proportional hazard regression analysis) is negative, the corresponding odds ratio (hazard ratio) will be smaller than 1. Conversely, if the beta coefficient is positive, the corresponding odds ratio will exceed 1.

Laparoscopic surgery. The beta coefficient is negative, resulting in an odds ratio smaller than 1. The odds ratio (laparoscopic versus open surgery) of having a postoperative infection is 0.59 (antilogarithm of −0.535). Again, interpretation of the coding is critical. In this example, laparoscopic appendectomy was coded as 1 (and open as 0). So, the odds of having a postoperative infection after laparoscopic surgery is only 0.59 times the odds of experiencing an infection after open surgery. If open appendectomy were coded as 1, the beta coefficient would be +0.535, and the corresponding odds ratio 1.69 (=1/.59), meaning that the odds of having a postoperative infection after open surgery are 1.69 times those of laparoscopic surgery. The change in coding would not result in a change of the p value.

COPD. The beta coefficient is positive, so the corresponding odds ratio is greater than 1. Patients with COPD are 1.11 times more likely to experience postoperative infections than patients without COPD. Again, if the coding were reversed (absence of COPD = 1, presence of COPD = 0), the beta coefficient would become negative (without a change in the absolute value: −0.10), and the odds ratio would be 0.9 (=1/1.11) and would be interpreted as the odds for non-COPD patients relative to patients with COPD. The 95% confidence interval crosses the value of the null hypothesis (for ratios = 1), resulting in a nonsignificant p value. Also, the beta coefficient is smaller than twice the standard error, which is indicative that the results are not significant.

How to assess the performance of statistical models

As described previously, statistical models are frequently used to adjust for confounding factors that could explain variability in outcomes. For example, two important factors that contribute to population variability in height are gender and age. When analyzing height as a function

of diet, one would naturally adjust for both gender and age.

Statistical models can also be used in an attempt to describe an outcome in terms of the factors that influence it. These types of models are called predictive models and their performance as such needs to be assessed. For linear regression models, a common assessment tool is the model R^2, which is the percent of variability in the outcomes that is jointly explained by the predictor variables in the model. The higher the R^2, the better the prediction of the statistical model. For example, the R^2 for a model of height (the outcomes) as a function of age among young people 5 to 17 years old might be 0.53, meaning that 53% of the variability of height is explained by age. Adding gender to the model might increase the R^2 to 0.61. Adding factors related to diet might further increase the R^2 to 0.73. Seventy-three percent of the variability in height is now explained by the predictor variables age, gender, and diet. The resultant model would still leave 27% of the variability unexplained. The residual variability is from factors that have not been included in the model or random variation.

For a dichotomous outcome analyzed by means of logistic regression, the situation is somewhat more complex. A predictive linear regression model attempts to estimate the actual outcomes for each individual, and the logistic regression model produces an estimate of the probability of experiencing the outcomes for each individual. The most common mechanism for describing how well a logistic regression model predicts the probability of experiencing the outcomes is the c-index, which is a concordance measure. For each pair of patients in which one experiences the outcomes (the case) and the other does not (the control), the pair is concordant if the estimated probability of the outcomes is higher for the case than for the control. The c-index is then the number of concordant case-control pairs divided by the total number of such pairs. A c-index close to 50% implies that the model cannot discriminate between cases and controls; a c-index close to 100% would indicate that, given the characteristics in the model, the estimated probabilities for cases are generally higher than those for controls. It is useful to view the c-index as the area under the receiver operating characteristics (ROC) curve, which is a summary curve portraying how well the estimated probabilities each separate the case population from the control population.

What statistical test should be used?

The choice of the appropriate statistical test primarily depends on several factors, including type of outcomes (continuous, categorical, or time to event), whether the data are paired (clustered) or unpaired, the number and type of risk factors or covariates being analyzed, and the assumed distribution of the data.

Outcomes

As discussed previously, the most frequently encountered end points in medical literature are continuous, categorical, or time to event. Although continuous outcomes cover a range (eg, tumor shrinkage measured in millimeters after neoadjuvant chemotherapy), categorical outcomes have only certain possible values (eg, dichotomous outcomes, such as response to chemotherapy, presence of metastases, and so forth). The time to event outcome is frequently used in surgical oncology, where studies evaluate the time after tumor resection to relapse or death.

Paired versus unpaired data

A paired (or clustered) data analysis should be used in the following situations:

1. If two or more measurements are done in the same subjects (eg, before and after an intervention)[31] (this would be a repeated measures design).
2. If we have matched pairs or clusters of subjects (eg, if, for each patient in a certain sample, another patient or patients with similar characteristics such as age, gender, race, and so forth has been assigned in a comparison sample, or if there are natural clusters such as members within families or patients within the same hospital).[31]

A paired test should be used to account for the lack of independence in the data and to obtain accurate p values.

Normal versus non-normal data distributions

Most statistical tests are either parametric (relying on a specified data distribution) or nonparametric (not relying on a specified data distribution). Many parametric tests rely on the normal distribution, which can be represented by the symmetric, bell-shaped Gaussian distribution curve. Parametric tests are more powerful than nonparametric tests if the underlying assumption about specific data distribution is met (Table 7).

It is often difficult to decide whether or not a parametric test should be selected. A nonparametric test is generally preferred in the following situations: If the

Table 7. Most Commonly Used Statistical Tests in Medicine

Parametric test	Corresponding nonparametric test	Null hypothesis to be tested	Example
Continuous outcome			
Two sample (unpaired) *t* test	Mann-Whitney-U test	Difference between means* from two independent (unpaired) samples is 0.	To compare the mean* values of gamma glutamyl transpeptidase between independent (unpaired) samples of patients with and without history of alcohol abuse.
One sample (paired) *t* test	Wilcoxon matched pair test	Difference between two measurements on the same sample (eg, before, after) or in matched patients is 0.	To compare the mean* tumor size in one sample of patients with rectal cancer before and after neoadjuvant radiotherapy.
One way analysis of variance (ANOVA)	Kruskal-Wallis test	Difference between means* from three or more unpaired/unmatched groups is 0.	To compare the average number of sampled lymph nodes in three unpaired/unmatched groups of patients undergoing different resection methods of pancreaticoduodenectomy for pancreatic cancer.
Repeated-measures analysis, including repeated measures ANOVA	Friedman test (analogue of repeated measures ANOVA)	Difference between trajectory from three or more paired/matched groups is 0.	To compare the trajectory of CEA values with respect to race (Caucasian, African American, others) of colorectal cancer patients at several different time points after resection.
Multiple linear regression analysis	Regression on ranks, other less frequently used tests	There is no association between a predictor variable and the continuous outcome after adjusting for potential confounding factors.	To evaluate whether age is independently associated with length of hospital stay in breast cancer patient undergoing lumpectomy after adjusting for race, socioeconomic status, comorbidities, tumor stage, and so forth.
Dichotomous/categoric outcome			
Chi-square test or Fisher's exact test†		Difference between proportions of the outcome from two (or more) independent/unmatched samples is 0.	To compare the proportion of successful endoscopic retrograde cholangiopancreatography in unmatched samples of elderly versus young patients with choledocholithiasis.
McNemar's test		Probability of the outcome is not more likely in one setting versus another (eg, pre or post or with one therapy versus another).	To compare the likelihood of response to proton pump inhibitors versus histamine antagonists for gastroesophageal reflux disease in matched patient samples.
Multiple logistic regression analyses		There is no association between the predictor variable and the categoric outcome after adjusting for potential confounding factors.	To assess the independent (risk-adjusted) impact of length of postoperative immobility on the occurrence of pulmonary embolisms after adjusting for age, race, socioeconomic status, comorbidities, and so forth.
Time to event outcome			
Log rank test (based on assumption that the relative hazard does not change over time)		Average hazards in the two groups are equal.	To compare the survival curves of colorectal cancer patients with liver metastases undergoing radiofrequency ablation versus cryotherapy.
Cox proportional hazard regression analysis (semi-parametric model, based on the assumption that the hazard ratio [hazard rate of group 1 divided by hazard rate of group 2] is constant over time).[32]		The relative hazard of two patient samples is one after adjusting for potential confounding factors.	To assess whether patients with cancer of the parathyroid gland with elevated postoperative parathyroid hormone have a shorter overall survival compared with patients with normal parathyroid hormone after adjusting for potential confounding variables such as age, gender, race, socioeconomic status, tumor size, grading, staging, and so forth.

*Non-parametric statistics test for equality of *medians* rather than *means*.
†For small sample sizes (< 20 patients), the chi-square test does not provide accurate results and Fisher's exact test must be used. For large samples, both Fisher's exact test and chi-square test yield similar results.

outcome is a score, eg, trauma score, comorbidity score, and so forth; if many outliers are present; and if the distribution of the data is clearly non-Gaussian. If it is unclear whether or not the data allow analysis by a parametric test, it is usually better to use a nonparametric test because the latter will yield slightly more conservative p values.

In conclusion, basic knowledge about statistical computations in medical literature is invaluable for critical assessment of scientific findings and their implementation in clinical practice. The learning curve for appropriate interpretation of biostatistics is steep and the process highly iterative. This article only scratches the surface of statistics in medicine. Some topics have been entirely omitted. Nonetheless, we hope that this *vade mecum* will provide a useful resource for surgeons and other physicians and will be a stimulus to enhance their ability to interpret statistical analyses.

Author Contributions

Study conception and design: Guller, DeLong
Drafting of manuscript: Guller
Critical revision: DeLong
Statistical expertise: Guller, DeLong

Acknowledgment: We thank Linda and Jonathan McCall for carefully reading the article and making many valuable suggestions; and Harvey Motulskey, whose book, *Intuitive Biostatistics,* inspired us as we drafted this article.

REFERENCES

1. Altman DG. Statistics in medical journals: developments in the 1980s. Stat Med 1991;10:1897–1913.
2. Hayden GF. Biostatistical trends in pediatrics: implications for the future. Pediatrics 1983;72:84–87.
3. Rothman KJ. A show of confidence. N Engl J Med 1978;299: 1362–1363.
4. How to read clinical journals: I. why to read them and how to start reading them critically. Can Med Assoc J 1981;124:555–558.
5. Evidence-Based Medicine Working Group. Evidence-based medicine. A new approach to teaching the practice of medicine. JAMA 1992;268:2420–2425.
6. Guyatt GH, Rennie D. Users' guides to the medical literature. JAMA 1993;270:2096–2097.
7. Berwick DM, Fineberg HV, Weinstein MC. When doctors meet numbers. Am J Med 1981;71:991–998.
8. Friedman SB, Phillips S. What's the difference? Pediatric residents and their inaccurate concepts regarding statistics. Pediatrics 1981;68:644–646.
9. Brown GW. Standard deviation, standard error. Which 'standard' should we use? Am J Dis Child 1982;136:937–941.
10. Jaeschke R, Guyatt G, Shannon H, et al. Basic statistics for clinicians: 3. Assessing the effects of treatment: measures of association. CMAJ 1995;152:351–357.
11. Munakata A, Nakaji S, Takami H, et al. Epidemiological evaluation of colonic diverticulosis and dietary fiber in Japan. Tohoku J Exp Med 1993;171:145–151.
12. Brodribb AJ, Humphreys DM. Diverticular disease: three studies. Part I—Relation to other disorders and fibre intake. Br Med J 1976;1:424–425.
13. Aldoori WH, Giovannucci EL, Rockett HR, et al. A prospective study of dietary fiber types and symptomatic diverticular disease in men. J Nutr 1998;128:714–719.
14. Kohler L, Sauerland S, Neugebauer E. Diagnosis and treatment of diverticular disease: results of a consensus development conference. The Scientific Committee of the European Association for Endoscopic Surgery. Surg Endosc 1999;13:430–436.
15. Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. N Engl J Med 1988;318:1728–1733.
16. Bland JM, Altman DG. Statistics notes. The odds ratio. BMJ 2000;320:1468.
17. Altman DG, Gore SM, Gardner MJ, Pocock SJ. Statistical guidelines for contributors to medical journals. Br Med J (Clin Res Ed) 1983;286:1489–1493.
18. Katz MH. Interpreting the analysis. Multivariable analysis. A practical guide for clinicians. Cambridge: Cambridge University Press; 1999:133.
19. Hanson MW, Feldman JM, Beam CA, et al. Iodine 131-labeled metaiodobenzylguanidine scintigraphy and biochemical analyses in suspected pheochromocytoma. Arch Intern Med 1991; 151:1397–1402.
20. Lenders JW, Pacak K, Walther MM, et al. Biochemical diagnosis of pheochromocytoma: which test is best? JAMA 2002;287: 1427–1434.
21. Berwick DM. Experimental power: the other side of the coin. Pediatrics 1980;65:1043–1045.
22. Freiman JA, Chalmers TC, Smith H Jr, Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. Survey of 71 "negative" trials. N Engl J Med 1978;299:690–694.
23. Perneger TV. What's wrong with Bonferroni adjustments. BMJ 1998;316:1236–1238.
24. Simon R, Altman DG. Statistical aspects of prognostic factor studies in oncology. Br J Cancer 1994;69:979–985.
25. Altman DG, Schulz KF, Moher D, et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Ann Intern Med 2001;134:663–694.
26. Pocock SJ, Hughes MD, Lee RJ. Statistical problems in the reporting of clinical trials. A survey of three medical journals. N Engl J Med 1987;317:426–432.
27. Vogel I, Francksen H, Soeth E, et al. The carcinoembryonic antigen and its prognostic impact on immunocytologically detected intraperitoneal colorectal cancer cells. Am J Surg 2001; 181:188–193.
28. Schott A, Vogel I, Krueger U, et al. Isolated tumor cells are frequently detectable in the peritoneal cavity of gastric and colorectal cancer patients and serve as a new prognostic marker. Ann Surg 1998;227:372–379.
29. Lindemann F, Schlimok G, Dirschedl P, et al. Prognostic sig-

nificance of micrometastatic tumor cells in bone marrow of colorectal cancer patients. Lancet 1992;340:685–689.
30. Guyatt G, Jaeschke R, Heddle N, et al. Basic statistics for clinicians: 1. Hypothesis testing. CMAJ 1995;152:27–32.
31. O'Brien PC, Shampo MA. Statistics for clinicians. 5. One sample of paired observations (paired t test). Mayo Clin Proc 1981; 56:324–326.
32. Peto R, Pike MC, Armitage P, et al. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. Br J Cancer 1976;34: 585–612.
33. Bland JM, Altman DG. One and two sided tests of significance. BMJ 1994;309:248.
34. O'Brien PC, Shampo MA. Statistics for clinicians. 8. Comparing two proportions: the relative deviate test and chi-square equivalent. Mayo Clin Proc 1981;56:513–515.
35. Friedman LM, Furburg CD, DeMets, DL. Sample size. In: Fundamentals of clinical trials. 3rd ed. New York: Springer Verlag; 1998, 94–129.
36. Peto R, Pike MC, Armitage P, et al. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. Analysis and examples. Br J Cancer 1977;35:1–39.
37. Schwartz WB, Mendelson DN. Hospital cost containment in the 1980s. Hard lessons learned and prospects for the 1990s. N Engl J Med 1991;324:1037–1042.
38. Fine MJ, Pratt HM, Obrosky DS, et al. Relation between length of hospital stay and costs of care for patients with community-acquired pneumonia. Am J Med 2000;109:378–385.
39. Gardner MJ, Altman DG. Estimating with confidence. Br Med J (Clin Res Ed) 1988;296:1210–1211.
40. Ma L, Chow JY, Cho CH. Effects of cigarette smoking on gastric ulcer formation and healing: possible mechanisms of action. J Clin Gastroenterol 1998;27(Suppl 1):S80–86.
41. Svanes C. Trends in perforated peptic ulcer: incidence, etiology, treatment, and prognosis. World J Surg 2000;24:277–283.
42. Svanes C, Soreide JA, Skarstein A, et al. Smoking and ulcer perforation. Gut 1997;41:177–180.
43. Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. BMJ 1995;310:170.
44. Bender R, Lange S. Adjusting for multiple testing—when and how? J Clin Epidemiol 2001;54:343–349.
45. Parker AB, Naylor CD. Subgroups, treatment effects, and baseline risks: some lessons from major cardiovascular trials. Am Heart J 2000;139:952–961.
46. O'Brien PC, Shampo MA. Statistical considerations for performing multiple tests in a single experiment. 1. Introduction. Mayo Clin Proc 1988;63:813–815.
47. Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. JAMA 1991;266:93–98.
48. Lee KL, McNeer JF, Starmer CF, et al. Clinical judgment and statistics. Lessons from a simulated randomized trial in coronary artery disease. Circulation 1980;61:508–515.
49. Rothman KJ. No adjustments are needed for multiple comparisons. Epidemiology 1990;1:43–46.
50. Stewart LA, Parmar MK. Bias in the analysis and reporting of randomized controlled trials. Int J Technol Assess Health Care 1996;12:264–275.
51. Duerr RH. The genetics of inflammatory bowel disease. Gastroenterol Clin North Am 2002;31:63–76.
52. Karban A, Eliakim R, Brant SR. Genetics of inflammatory bowel disease. Isr Med Assoc J 2002;4:798–802.
53. Yang H, Rotter JI. The genetic background of inflammatory bowel disease. Hepatogastroenterology 2000;47:5–14.
54. Satsangi J, Parkes M, Louis E, et al. Two stage genome-wide search in inflammatory bowel disease provides evidence for susceptibility loci on chromosomes 3, 7 and 12. Nat Genet 1996; 14:199–202.
55. Pocock SJ, Clayton TC, Altman DG. Survival plots of time-to-event outcomes in clinical trials: good practice and pitfalls. Lancet 2002;359:1686–1689.
56. Altman DG, Bland JM. Time to event (survival) data. BMJ 1998;317:468–469.
57. Mullner M, Matthews H, Altman DG. Reporting on statistical methods to adjust for confounding: a cross-sectional survey. Ann Intern Med 2002;136:122–126.
58. Ramstedt M. Per capita alcohol consumption and liver cirrhosis mortality in 14 European countries. Addiction 2001;96(Suppl 1):S19–33.
59. Singh GK, Hoyert DL. Social epidemiology of chronic liver disease and cirrhosis mortality in the United States, 1935–1997: trends and differentials by ethnicity, socioeconomic status, and alcohol consumption. Hum Biol 2000;72:801–820.
60. Green MS, Symons MJ. A comparison of the logistic risk function and the proportional hazards model in prospective epidemiologic studies. J Chronic Dis 1983;36:715–723.
61. Taulbee JD, Symons MJ. Sample size and duration for cohort studies of survival time with covariables. Biometrics 1983;39: 351–360.
62. Schoenfeld DA, Richter JR. Nomograms for calculating the number of patients needed for a clinical trial with survival as an endpoint. Biometrics 1982;38:163–170.
63. Schoenfeld DA. Sample-size formula for the proportional-hazards regression model. Biometrics 1983;39:499–503.
64. Makuch RW, Simon RM. Sample size requirements for comparing time-to-failure among k treatment groups. J Chronic Dis 1982;35:861–867.
65. Lachin JM. Introduction to sample size determination and power analysis for clinical trials. Control Clin Trials 1981;2:93–113.