Endovascular Flow Splitting

TO THE EDITOR: In the March issue of the Journal of Neurosurgery, Takahashi et al. (Takahashi JC, Murao K, Iihara K, et al: Successful “blind-alley” formation with bypass surgery for a partially thrombosed giant basilar artery tip aneurysm refractory to upper basilar artery obliteration. J Neurosurg 106:484–487, March 2007) reported the successful treatment of a giant, partially thrombosed aneurysm of the basilar artery (BA) tip region by direct clipping of the BA tip, left posterior communicating artery (PCoA), and P$_2$ segment.

Abstract

Partially thrombosed giant aneurysms that are located at the basilar artery (BA) bifurcation and are not amenable to clip application are among the most challenging lesions for neurosurgeons. They compress vital structures such as the brainstem and the thalamus, and the prognosis is extremely poor when they are left untreated. Although obliteration of the upper BA is a promising approach for these aneurysms, some lesions are refractory to this treatment, and effective additional strategies have not been clearly established. The authors report a case treated by placement of clips in the unilateral posterior cerebral artery (PCA) and posterior communicating artery as well as by superficial temporal artery–PCA bypass after unsuccessful upper BA obliteration. Complete thrombosis and dramatic shrinkage of the aneurysm were obtained.

Vessel occlusion transformed the aneurysm in the dead end of a vascular “alley.” A distal bypass furnished the left PCA, which was occluded at the P$_2$ segment. Such a risky operation was performed in 2 steps, first by direct clipping of the BA tip, which was ineffective in reducing lesion growth, and second by clipping the PCoA and P$_2$ segment combined with bypass surgery on the distal PCA. Direct coiling of the patent portion of the aneurysm was judged infeasible or feasible but ineffective on the thrombosed portion of the aneurysm. The happy end of the story—that is, complete anatomical and clinical cure of the patient—compels us to congratulate the authors.

On the other hand, at least 2 different endovascular treatments, less invasive than direct surgery, would have been possible. First, simultaneous coil obliteration of the patent portion of the aneurysm and the P$_1$-P$_2$-PCoA junction would have produced the same anatomical result on flow splitting, with the added benefit of obliteration of the aneurysm. Obliteration of the parent vessel guarantees anatomical cure of the lesion, regardless of the presence of a thrombosed portion. As clearly shown in their Figs. 1B and 3A and C, the aneurysm neck does not exactly involve the BA tip, but rather the left P$_1$-P$_2$-PCoA junction. We and others have applied similar treatments in similar cases of aneurysms, with subsequent complete anatomical and clinical cure (Fig. 1). Second, coiling of the aneurysm and contemporaneous positioning of a stent on the P$_1$-P$_2$ segment might allow long-term anatomical reconstitution of the neck. Note, however, that the concept that stents allow better long-term results in large aneurysms by improving neck healing, a common theory in the European neuroradiological community, has not been proven and thus remains an opinion.

As regards perfusion of distal PCA territories, distal bypass is a possible solution. In our experience, endovascular occlusion of the P$_2$ segment in 8 consecutive patients did not cause any clinically evident distal infarctions because

Fig. 1. Angiograms obtained in a 47-year-old woman who experienced the sudden onset of left cranial nerve III palsy. A: Left vertebral artery angiogram obtained in the arterial phase, oblique view, showing a large aneurysm of the P$_1$-P$_2$-PCoA junction. Arrows indicate the left PCA arising from the aneurysm wall. B: Platinum coil occlusion of the aneurysm and the P$_1$-P$_2$-PCoA junction. C: Left internal carotid artery angiogram obtained in the late arterial phase, lateral view, demonstrating retrograde filling of the peripheral PCA (arrows) through leptomeningeal collateral vessels. The patient had no clinical deficits or ischemic lesions on postoperative magnetic resonance imaging.
of the rich collateral circulation; other authors have report-
ed similar results.¹

Mauro Bergui, M.D.
Gianni B. Bradac, M.D.
S. G. Battista Hospital
Turin, Italy

Reference

1. Ciceri EF, Klucznik RP, Grossman RG, Rose JE, Mawad ME: Aneurysms of the posterior cerebral artery: classification and en-

RESPONSE: We thank Drs. Bergui and Bradac for their comments. First, we must say that our paper focused on ways of dealing with the partially thrombosed BA tip giant aneurysm when the classic obliteration technique involving occlusion of the upper BA results in failure. Although BA obliteration has long been performed with a measure of success in cases of unclippable giant aneurysms at the BA tip, this strategy is ineffective in some cases, and no de-
tailed technical reports on the additional surgery are available. We intended to show that another option exists even in cases previously considered hopeless. Needless to say, the treatment in our case was not an intentional 2-stage operation.

As to the general management of BA tip giant aneu-
rysms, Drs. Bergui and Bradac have proposed 2 promising strategies. One is coil embolization of the aneurysm lumen with simultaneous occlusion of the PCA. Indeed, they have shown a successfully treated case of a large aneurysm arising from the left P₂-P₃ junction. However, our case differs anatomically. We do not agree with their comment that “the aneurysm neck does not exactly involve the BA tip.” Although it is true that the patent portion of the aneurysm lumen leans leftward and involves the left P₂-P₃-PCoA junction, 3D computed tomography angiography clearly demonstrates that the BA tip is also affected (Fig. 2B in our paper). Furthermore, in the operative field we observed involve-
ment of the BA tip in the extremely wide neck. If the parent artery was occluded with coils, the vital perforating vessels that arise from the BA tip area could have been sacrificed and serious complications would have occurred.

Another strategy proposed by Drs. Bergui and Bradac is the use of an intracranial vascular stent. Although this approach appears promising, these devices are, unfortunately, not yet available in our country. We hope such new technol-
ologies and devices will help us to treat complicated giant aneurysms that cannot be cured at present.

We simultaneously performed a superficial temporal ar-
tery–PCA bypass and obliteration. As Drs. Bergui and Bradac have commented, collateral networks between the PCA and other arteries are developed to a considerable extent in many cases. However, we selected a reliable method to prevent infarction in the occipital lobe because our patient had left piosis at the time of surgery and would have been extremely distressed if homonymous hemian-
opia were to occur as a result of occipital infarction. (DOI: 10.3171/JNS/2008/108/6/1255)

Jun C. Takehashi, M.D.
Susumu Miyamoto, M.D.
National Cardiovascular Center
Osaka, Japan

Biostatistics With Neurosurgical Importance

TO THE EDITOR: All investigators as well as editors strive to use, evaluate, and analyze data in the most appropriate way. In this letter I briefly depict some pitfalls of using statistical analyses. Problems can arise when researchers use statistical tools such as regression, correlation, recall bias, sampling distribution, sample size, type I and II errors, end point, and so forth. To demonstrate these problems, 5 articles (Howington JU, Kutz SC, Wilding GE, et al: Cocaine use as a predictor of outcome in aneurysmal subarachnoid hemorrhage. J Neurosurg 99:271–275, August, 2003;¹¹ Pal-
el. J Neurosurg 98:1217–1221, June, 2003¹⁵) published in past issues of the Journal of Neurosurgery were used purely for the purposes of illustration and general discussion. Working through the examples will help to highlight the is-

issues involved plus demonstrate how to potentially solve the problems.

Generally speaking, there have been 2 commonly used methods in medical statistics for describing the relationship between 2 sets of variables (clinical factors): 1) correla-
tion study; and 2) regression analysis. Although both can be used for prediction,¹⁵¹⁶ regression analysis is certainly the more preferred and more definitive test. There are differ-
ences between these 2 methods. Correlation is a measure of the association between 2 variables; however, no causal relationship is implied. With regard to determining corre-
lation, the Pearson correlation coefficient has been used to indicate the extent to which 2 variables change with one another in a linear fashion. Conversely, the Spearman rank correlation test, such as that used in the study by Yamamoto et al.,¹⁶ is a nonparametric measure of the correlation be-
tween 2 variables with no assumptions about the frequency distribution of the variables. Regression analysis is used to determine whether there is any significant statistical relation-
ship between 2 independent variables. It has been noted that there are different types of regressions, such as logistic regression (supposedly used in the study by Tyler-Kabara et al.,¹⁷ but with marked violation of the statistical defini-
tion), multilinear regression, and Cox regression.¹⁵,¹⁶,¹⁹,¹⁸,₂⁵,₃₀

Statistical methods require the assumption of the null hypothesis as true. All such assumptions must be strict-

ly followed. Nevertheless, some authors have violated the statistical assumptions of a study.¹⁷,¹⁹,¹₂,¹₄,₂₆–₂₉ The clinical researcher must consider the hypothesis of the statistical model used. For instance, in a Cox regression model, which is most commonly used for survival analysis,¹ the statistical assumption is a reduction in the proportional hazard. Fur-
thermore, such an assumption is carried across time, regard-
less of the first 48 hours after cerebral aneurysm bleed-
ing or any longer period of time. Generally speaking, on regression analysis of survival data, the ratio of the hazards comparing different exposure groups remains constant over time. This is called the proportional hazards assumption. Moreover, in Howington and colleagues’ study on cocaine use, there are survivals without cerebral aneurysm rebleeding at 5 days, 1 month, or 3 years—examples across time. Thus, the statistical assumption, which is embedded in the Cox regression model used for survival analysis, is across time.

Among all types of regressions, the logistic regression merits our attention too. In a study of trigeminal neuralgia (TN), Tyler-Kabara et al. performed a regression analysis to handle their data, and at first glance it appears to be a logistic regression (see Reporting Statistics Clearly and Properly in Medicine). However, they used outcome categories of “cure” (excellent pain relief), “partial” (partial or good pain relief), or “none” (no pain relief), rather than a quantitative measurement. Technically, one of the characteristics in logistic regression is the appropriateness of binary outcome variables. However, according to Tables 1 and 2 in the Tyler-Kabara et al. study, there are 3 categories (cure, partial, and none), instead of 2 (a binary of “yes” [cure] and “no” [no cure]). The authors appear to be violating the basic assumption of distribution in a logistic regression. In addition, they apparently feel the need to have an extension of the logistic regression, which is acceptable when the outcome variable is ordered categorical. It would be easier for readers faithful to the definition of logistic regression if the authors would clearly explain such an extension. With the aforementioned precaution in mind, I do not attempt to teach a very wide range of clinical research/biostatistical concepts with examples drawn from existing published textbooks. On the contrary, I humbly aim to highlight the pitfalls of statistical analyses, as outlined in the Results and Discussion. By working through the examples, one can easily highlight the issues involved and offer possible solutions.

The organization of this article starts with the diagnosis period and its influence on the treatments assigned. The following sections cover different clinical stages of the same samples (patients), the interaction between 2 variables (clinical factors), sampling distribution, sample size, type I and II errors, recall bias, end point, and so forth. I will also address how the significance of a relation between variables depends on sample size. In this article I also discuss the problems of selecting statistical models and reporting medical statistics properly.

Materials and Methods

The clinical research material used in the present study was gathered from 5 articles published in the Journal of Neurosurgery. These studies cover chordoma, TN, cocaine use, carotid endarterectomy (CEA), and chronic subdural hematoma (CSDH). The analysis of these articles is for purely academic purposes, and it is not my intention to criticize the work of any of the authors. All data were used in accordance with the spirit and principle of the Health Insurance Portability and Accountability Act regulations (1996). All data were obtained from the published studies, and their sources are specified.

In the current analysis, the medical statistic principles and methodology used are significant, and a systematic approach was applied as much as possible, both clinically and statistically. Explanations of statistical tests (which can be found in any textbook) are not excessively repeated in this article, and a minimum of statistical formulas are applied here. The mainstream philosophy of science is respected and followed in this study based on 1) the viewpoint that knowledge is justified, 2) real persuasion, and 3) the epistemology founded on such a view predominates. Based on such an understanding, the forthcoming result and discussion are in order.

Results and Discussion

The Diagnostic Period and its Influence on Assigned Treatments

As patients age some have a recurrence of neuralgia, whereas others experience a progression from atypical to typical TN. Conversely, a few patients pass from the stage of chronic to recurrent subdural hematoma. As patients age the amount of time from the initial establishment of a diagnosis increases too.

Patients at high risk for early tumor recurrence can enter trials featuring aggressive surgical treatment in which multiple and extensive surgical approaches are followed by proton beam radiation therapy. Pallini and colleagues investigated the relationship between the biological behavior of chordoma, which was determined based on the interval of the lesion’s recurrence and the lesion volume doubling time, and several pathological and molecular features. In their study there were 14 men and 12 women who ranged in age from 17 to 80 years (mean 50.3 years). There were no data about the age of the patients when diagnoses were established; that is, no diagnostic period was clearly indicated in their report. Any tumor recurrence was acknowledged from the time of its diagnosis, instead of when surgical treatment was begun. Early or late diagnosis should have been shown and defined.

There were at least 5 treatment options that could be assigned to any new patient with chordoma. These options consisted of total removal without radical resection, subtotal removal without radical resection, radical resection, a transsphenoidal biopsy procedure before craniotomy, and no surgery. Given the categories of early and late diagnostic periods and assigned treatment, there were 10 (2 × 5) different possible combinations.

Similarly, among studies on TN, age is considered an important factor. Patients in the atypical TN group—both in the Tyler-Kabara et al. study and in the general population—were younger before they entered the stage of typical TN. Conversely, the recurrence of TN was often misdiagnosed as a new occurrence in new patients. This mistake is not only confusing but also misleading (see Different Clinical Stages in the Same Patient [Sample]).

Similarly, in the Yamamoto et al. study of CSDH, the number of patients in the recurrent group was 11, whereas that in the nonrecurrent group was 94 (ratio 11:94). For 105 patients with CSDH, there was no recorded diagnostic period, which was needed to calculate the time interval involved. For any recurrence of CSDH, one must consider
the time when the diagnosis was established, instead of when surgical treatment commenced. From an epidemiological viewpoint, the diagnostic period should have been considered a confounding factor. To statistically adjust the confounding factor, one can easily rectify the confounding degree in a multivariate analysis, just as was done in the study by Yamamoto et al.

Different Clinical Stages in the Same Patient (Sample)

Whenever clinical research is a crossover trial, each sample (patient) at a different clinical stage manifests an “independent” clinical presentation. In reality, at each stage, the clinical picture comes from the same sample (patient). There is a time relationship in terms of the data collected at each clinical stage. Among the 5 studies under discussion, the clinical stages of TN are preoperation, postoperation, and before and after reoperation. There are only stages of recurrence and nonrecurrence for CSDH. If the researcher uses the chi-square test for a statistical inference, he or she violates the basic assumption of this test because the data are supposed to be independent of each other between samples. Hence, the McNemar test, instead of the chi-square test, is a more suitable tool.

It is noted that the Yates correction for the chi-square test is most important, especially when one moves from a continuous to a categorical scale (see Reporting Statistics Clearly and Properly in Medicine). The Yates correction factor is used in small samples for decreasing the type I error; that is, the null hypothesis is less often rejected. We can incorrectly reject the null or test hypothesis; for example, we can conclude that a medication decreases the mortality rate when in fact it has no effect. This mistake is known as a type I error (see A Small Sample Size and its Effect on Type I and II Error Rates). The Yates correction decreases a chi value but also increases the probability estimate.

Interaction Between 2 Independent Variables (Clinical Factors)

In the study of cocaine use and aneurysm bleeding, there was an interaction between 2 variables: cocaine use and aneurysm bleeding (A) and the assigned treatment (B). Therefore, the possibility of an interaction between A and B, the mathematical product of the 2, must be considered.

The nature of an interaction between variables is always relevant. In the Shimada et al. study on CEA, a high index of diagnostic suspicion depended on the percentage of patients with severe carotid artery disease (CAD). The percentage of patients with CAD increased stepwise as the number of major risk factors increased. The percentage also depended on whether patients with severe CAD had more major risk factors than those with mild CAD. Hence, it is worthwhile to consider the interaction between CAD and peripheral artery disease (PAD). The following model, in which diabetes mellitus is indicated by DM, can then be developed: age + DM + PAD + (DM × PAD). Clinically, an interaction, such as DM × PAD, is inevitable, and one should be alert to it. Biostatistically, the way to eliminate the interaction is to have both forward and stepwise selections in the multivariate analysis, although this topic is not the focus of the current paper.

Sampling Distribution of Patients

In their study on TN, Tyler-Kabara et al. used the median and the mean of the same data (see Table 2), which required an additional explanation, as the median is for nonnormal distribution and the mean for normal distribution. To compare 2 sets of continuous data in a normal distribution, one can use the Student t-test. If the 2 sets of continuous data are in a nonnormal distribution, it is better to use other tests. To use both the median and mean of the same data, the authors must have a good reason. Without any explanation, however, they may frustrate readers.

Statistically, it is important to ascertain whether a sample is or is not in a normal distribution. Such a step is needed not only in reading the study on TN, but also in looking at any other reports. In Table 3 in the report on TN, if the data had been plotted on a graph and shown to be a nonnormal distribution, the graph would have been positively skewed. In this case, it would be appropriate to use a non-parametric test for the data set.

In Table 3, one has to imagine whether the chi-square test was used to generate the statistical test values, as it was never specified in the article. In other studies—those on chordoma, CEA, and CSDH—the statistical methods were clearly indicated by the authors.

An additional issue is as follows: Table 3 is entitled “Clinical history in typical and atypical TN.” Each clinical history variable should have been based on the ratio (“TN Diagnosis [%]”) of the number of diagnoses denoted in the table’s title; it is not supposed to be any “number of patients.” The information presented in Table 3 is so confusing. There is an obvious difference between the number of clinical histories and the number of patients. These factors should have been clarified.

Without clarification, the table will frustrate readers even more. According to the authors of the TN study, “previously collected data relating to complications and outcomes were combined with data collected through phone calls to patients to yield the information in this manuscript” and in Table 3. If one uses the number of operations as the sample, how can one know that each patient with multiple operations could actually recall the exact difference in their clinical history including, but not limited to, the complications of each different operation?

If we regard the number of patients as the sample in Table 3, an additional problem arises in determining data. For patients with multiple operations, with regard to data (%) based on telephone calls, the patient responses are mainly based on which operation(s)? There may have been confusion between 2 or among 3 or more different operations for a single patient. Hence, which patient responses to the telephone calls were based on which operation(s) is still unclear. In contrast, if the data were obtained from patients who had undergone a single operation and long-term follow-up, there then would have been no confusion regarding the responses. Thus, there would be no need for a variable such as prior microvascular decompression (see Recall Bias).

Hence, the proper concept of sampling distribution—for instance, in a one-sample t-test—has a null hypothesis.
something like “the sample mean consistent with the sampling distribution of all sample means with the same underlying population (that has the population point estimate as the true mean, \( \mu \)).” Therefore, if one fails to reject the null hypothesis—that is, your probability value is > 0.05—then one concludes that the sample mean obtained is consistent with the population mean, \( \mu \). This reasoning leads one to conclude that there is no artifact of sampling; that is, there is no evidence in a sample that there was any difference from the underlying population. To get a better idea of what all of these explanations mean, a clearer picture should have been painted with regard to the number of patients versus the number of operations, as discussed earlier. The aforementioned issues merit serious attention to determine whether a sample represents the patients or the operations (see Reporting Statistics Clearly and Properly in Medicine).

**Recall Bias**

Recall bias has been only briefly mentioned in a preceding section. Nevertheless, there seems to be another troublesome issue in Table 4 in the study of TN.\(^{28}\) In the case of patients who underwent multiple operations for TN, one must figure out 2 more problems. First, after which operation did the postoperative complication occur? Second, were all complications, regardless of their number, counted in Table 4? Furthermore, because the same sample (patient) was counted on more than one occasion (operation), then combining the probabilities would not engage the simple multiplication rule for the combined occurrence of independent events. Conversely, what about the complications forgotten by the patients? Hence, there is a recall bias. This factor is very important and could influence the validity of a study finding. Such a bias can occur when the patient’s memory or historical facts are based on the current state of disease or the social appeal of certain lifestyle factors.

**A Small Sample Size and its Effect on Type I and II Error Rates**

Note that the number of patients with recurrent subdural hematoma in the study of CSDH\(^{11}\) is 11, and in the study of chordoma\(^{23}\) it is 26. Such sample sizes cannot be considered large. When the sample size is small and a strict significance level (instead of 0.05, setting \( \alpha \leq 0.01 \)) has been set to minimize the type I error (the error of statistically rejecting a true null hypothesis), there may be an increase in the probability of a type II error. A type II error statistically accepts a false null hypothesis. The probability of a type II error can be considered when a wider confidence interval is associated with a small sample size, such as the aforementioned 11 patients. Results from a study with a small number of patients can suggest that a particular treatment could help patients with a certain disease (or a test could assist in the diagnosis of a certain disease). For more conclusive results, the study should be extended to a much larger patient population before accepting a therapy as the treatment of choice (or a method as a diagnostic tool with clinical value). The reliance of medical decisions on narrowing primary end points can inflate both type I and II error rates. Note also that, at times, it is inevitable for the primary end points to have been narrowed at least to a certain extent, such as dichotomizing a continuous variable into only 2 groups. In this circumstance, to address these concerns of potential type I and II error rates, a solution is offered. All the end points, especially the primary ones, should be as informative as possible. To avoid losing information during the transformation from a continuous variable into an artificially categorical one, the concept of information-preserving composite end points must be defined. Furthermore, when this type of composite end point would be useful should be established.

**The True Definition of an End Point**

One very hilarious thing about neuroscience writing is that one word might fit perfectly—“end point” is one, and “capricious” is another. The latter word is a good description of how a Golgi stain permeates some neurons perfectly and completely, whereas other neurons remain unstained. In fact, “capricious” is not really appropriate in a biochemical sense, as the term means “random.” Outside of biochemistry, “capricious” has a connotation of carelessness or other negative implications. Similarly, “end point” in certain settings has no meaning apart from denoting the limit of something, whereas in other settings it can have a more colorful meaning. Conversely, confusion regarding the ends of an interval comes from the difference between a parenthesis and a bracket, or a \( < \) and a \( \leq \) symbol. When one of these symbols is used you are including the value at the end of the range. When another symbol is used, you are specifying that a particular value is not included. If an author states \( x < 15 \), \( x \) can never equal 15 but must be \( < 14.99999 \). In contrast, if \( x \approx 15 \), \( x \) can be as large as 15.0, but no larger. Hence, the term “end point” is confusing. In statistics, end points are referred to as “limits,” which implies that the values themselves are included as possible values (which they are). One can loosely define an end point as merely a measure of disease severity. It can be regarded as a standard in the literature. Nevertheless, different diseases have various measures of severity. For example, for sleep apnea, the measure of severity is the apnea-hypopnea index, which is the number of apnea and hypopnea events per hour. It is a rate, not a single point (the end point), and a continuous variable. In the literature of sleep medicine, the apnea-hypopnea index can hardly be referred to as any single point but a rate. As another example, according to an international multicenter randomized trial comparing standard pressure differential valves,\(^{11}\) Orbis Sigma valves and PS Medical Delta valves, in children with newly diagnosed hydrocephalus, there was no dissimilarity in the time to first shunt failure (power 80\%). To avoid observer bias, the primary end point, shunt failure, was defined by detailed clinical and radiological criteria. An “end point” does not mean the “measure” or “severity” of diseases. In fact, there are different end points in certain diseases, and thus we need to combine them and improve the informational content of categorical clinical trial end points.

**A Large Disparity Problem in the Sample Size**

Next to the interesting discussion on end points, “sample size” merits our consideration. In the recurrent group in the CSDH study\(^{21}\) the sample size of 11 patients is small, whereas its counterpart, the nonrecurrent group, consists of 94 patients. There is a large disparity problem in these sam-
ple sizes; therefore, it is important to use the Fisher exact test to generate the probability values pertaining to any chi-square statistic that is obtained. The Fisher exact test is very significant if the observed number in any cell is < 5. The appeal of this test is that the true probability value is always smaller than the calculated value; hence, the Fisher exact test is a conservative method of analysis. In using this test, there is strong evidence of a difference indicates a statistically significant difference. The inexact use of the Fisher exact test was reported in 6 major medical journals in 1989. Applying this test without specification as a one- or two-tailed test can misrepresent the statistical significance of data. Note the following example: The authors of 33 of 56 selected articles did not specify the use of a one- or two-tailed test, and in 12 (36%) of these articles the one-tailed test was used. Five (42%) of these 12 articles contained at least 1 table in which the standard significance level < 0.05 was no longer met when a two-tailed analysis was performed. If there were really 2 independent samples, each with 11 cases for a comparison, the chi-square test would be appropriate. Because of the small sample size and because not all the theoretical values were not less than 5, the Fisher exact test would have been the appropriate statistical method.

Significance of a Relationship Between Variables Depending on Sample Size

If there are very few observations, there are also few possible combinations of values of the variables. Nevertheless, the chance of taking these values as indications of a “strong relationship” is relatively high. The smaller the sample size in a research design, the more likely are erroneous results. Statistical tests performed on small sample sizes can reveal a relationship between 2 variables that does not even exist. In fairness, with regard to small samples, such as the 11 patients in the recurrent group of the CSDH study, the statistical aspect of a sample size is governed by the goal of the designed study. For instance, it is entirely possible to have a Stage 1 study in which the authors are trying to show any accuracy > 0.5. In such a circumstance, one need only show that the area under the curve of a receiver operating characteristic (ROC) analysis is statistically > 0.5, and this can be accomplished with a sample size of 11 nondiseased and 11 diseased patients, without violating any assumption of the ROC analysis.

Reporting Statistics Clearly and Properly in Medicine

The next puzzle in the study of TN relates to the statistical model. The positive and negative predictors as figured in that study are puzzling. The authors mentioned the use of independent variables to indicate whether the results of the statistical analysis were significant. They were testing with a regression model(s). Generally speaking, a positive predictive value is defined as the proportion of positive tests that identify diseased persons. A negative predictive value is the fraction of negative tests that correctly classify nondiseased persons. It is unclear what type of correlation was calculated in the study of TN. The reported outcomes are not quantitative but ordinal.

Based on the appearance of independent variables in the study on TN, readers have no choice but to apply the process of elimination to determine if simple or multiple regressions were used by the authors. Furthermore, in Tables 6 and 7 of that study, there are categories of “cure,” “partial,” and “none,” which represent a trichotomy instead of binary outcome variables. Binary outcome variables are supposed to be shown in a logistic regression (see last paragraph of Materials and Methods). If the statistical methods had been clearly indicated, the reader would not have to try so hard to figure them out. Generally speaking, when the outcome measure used in a logistic regression is a trichotomy, instead of a dichotomy, it is appropriate to use multinomial logistic regression. The original authors must establish the reference category in such a regression.

As to the trichotomy of outcomes (cure, partial, and none), as opposed to binary outcome variables, note the following points. By using the term “partial,” it is implied that results cannot be differentiated between improved and not improved (or in the context of the TN study, between “cure” and “none”). In the TN study, the validity of the statistical method used in arriving at the significance of such findings is most important.

Table 4 of the TN study is entitled “Complications associated with 2003 MVDs [microvascular decompressions] for typical and 672 for atypical TN.” The data in the table include only the number of complications instead of the number of patients. A similar situation occurs in Table 3, as previously discussed. The authors should specify whether the data indicate the number of operations or patients. Given that there is the variable of “death” in the last line of Table 4, one can conclude that the numbers indicate patients instead of operations, although the table title is misleading. The problem in Table 4 is its title: “Complications associated with 2003 MVDs for typical and 672 for atypical TN.” The abbreviation MVD refers to microvascular decompression and TN to trigeminal neuralgia; therefore, all reported numbers in the table should refer to complications of procedures and not the number of patients. Unfortunately, when “death” was listed as a complication of a procedure, it appears more appropriate to have a proper table title to clarify such a particular but irreversible complication.

Such an issue is similar to the confusing issue in Table 3, as previously discussed. Unlike in Table 4, however, the number of patients was quite small in Table 5. A minor increase in such a small number of patients by merely 1 or 2 individuals can alter statistical significance (see A Small Sample Size and the Effect on Type I and II Error Rates).

The data in Table 5 consisted of ordinal numbers. Hence, the chi-square test was no longer applicable. Instead, a Mann–Whitney U-test, a nonparametric method, was used. Nevertheless, among the postoperative data, it was unclear which instance of pain relief in patients who had undergone multiple operations should have been counted. The possibility of recall bias in the patients must be considered (see Recall Bias).

The data in Tables 6 and 7 can only be properly discussed after the source of the data in the preceding tables is clarified.

It is very important to report statistics clearly and properly in medicine. A statement such as the one in the study on TN—“All data were analyzed by a statistician (Y.F.C.), who then made graphs of the results”—is really suboptimal. It is not clear what kind of statistical methods
Conclusions

In summary, there is indeed a difference between clinical importance and statistical significance. Sometimes, results can surface by chance but still indicate a real difference. Of course, the difference can be so small as to render it unimportant for all practical purposes. With such an understanding, I used 5 clinical studies published in the Journal of Neurosurgery to demonstrate some statistical characteristics and some pitfalls of these research reports. I have emphasized regression, correlation, prediction, recall bias, sampling distribution, sample size, type I and II errors, end points, and more. Before reaching any significant inference in statistics, I hope all clinical researchers will pay attention to the nature of the data, the influence of sample size, the interaction between variables, and the statistical hypothesis and deal properly with the data by selecting appropriate statistical models and reporting medical statistics accurately.

Acknowledgments

I thank Dr. Evan Balaban, McGill University, Montreal, Canada, for his opinion and advice after the initial reading; Dr. Jaypaual Azariah, University of Madras, Guindy Campus, Chennai, India, for his comments and opinions; Dr. Richard Sulter, St. John’s University, Jamaica, New York, for his critical reading; and Dr. Rochelle E. Tractenberg, Georgetown University School of Medicine, Washington, DC, for her input in The True Definition of an End Point.

Bing Tang, M.D., M.P.H.
Research and Ethics
Danville, California

References

6. Faller H: [Sensitivity, specificity, positive and negative predictive value.] Rehabilitation (Stuttg) 44:44–49, 2005 (Ger)

17. Lo SK, Li IT, Tsou TS, See L: [Non-significant in Univariate but Significant in Multivariate analysis.] Changgeng Yi Xue Za Zhi 18:95–101, 1996 (Chn)
RESPONSE: Dr. Tang has reviewed 5 articles published in past issues of the Journal of Neurosurgery, including our paper on clival chordoma. Based on these articles, he describes some of the pitfalls of using statistical analyses. More specifically, Tang has mentioned our article on clival chordoma 7 times, with regard to the following issues: use of the Pearson correlation coefficient; indication of the diagnostic period(s) and time at diagnosis; acknowledgment of tumor recurrence; patient grouping relative to treatment options and time of diagnosis; clinical staging of the same patient (sample) and use of the chi-square test; indication of the statistical method used; and the small number of patients with chordoma (26 cases).

Pearson Correlation Coefficient

Tang noted that the relationship between 2 variables was analyzed in the articles on chordoma and CSDH by using the Pearson correlation coefficient and Spearman rank correlation test, respectively. In our study on chordomas, we plotted the tumor volume doubling time and proliferation index (Ki 67) of recurrent cases of typical chordoma in an x–y diagram and found that the 2 variables followed an exponential curve (see Fig. 5 and not Table 4 as erroneously indicated by Tang). We then transposed data in a semilogarithmic diagram and applied the Pearson linear correlation coefficient, which produced a p = 0.032. Tang criticized the use of the Spearman rank correlation test in the study on CSDH; however, he did not clearly state whether the Pearson correlation coefficient used in our study was correct or not. Given that Tang has included our article among those with statistical flaws, we would expect to know exactly what is supposed to be wrong with our methods or, in the event that our method is correct, we would like this concept to be easily understood by any reader.

Diagnostic Period(s) and Time at Diagnosis

Tang correctly points out that we investigated the relationship between the biological behavior of chordoma, which was determined according to the interval of tumor recurrence and the tumor doubling time, and several pathological and molecular features. He claims that both the diagnostic period and time at diagnosis are not indicated in our article; however, Table 1 shows the age of the patients at admission (second column), the interval (in months) between subsequent surgeries (fourth column), and whether each operation was performed at our institution or elsewhere (fourth column and footnotes). In patients undergoing their first operation at our institution, the age at diagnosis (histological diagnosis) coincides with the time that surgical treatment commenced. In patients in whom initial surgery was performed elsewhere, diagnostic periods are also shown in the fourth column of Table 1. For example, the patient in Case 1 was 55 years old at the time of admission, which coincides with both the time of his first transsphenoidal (TS) operation and the time of histological diagnosis. This patient then underwent 2 TS surgeries followed by radiotherapy. The intervals for the first and second tumor recurrence were 5 and 8 months, respectively. The patient in Case 22 was admitted at 46 years of age after 3 transoral operations performed elsewhere; the intervals for the first and second tumor recurrence were 44 and 28 months, respectively. In this case, the time at diagnosis was 46 years — 6 years (72 months), that is, 40 years. Tang also remarks that any tumor recurrence in our study was acknowledged from the time when the diagnosis was established instead of the time when surgical treatment was begun. However, the time of diagnosis coincides with the date of the first operation when a diagnosis of chordoma was established on histological examination.

Acknowledgment of Tumor Recurrence

Tumor recurrence was defined as “progression on clinical or neuroimaging assessment that required reoperation” (see Clinical Material and Methods in our article). In each patient, the dates between surgical procedures represented the end points to calculate the interval for tumor recurrence.

Patient Grouping Relative to Treatment Options

Tang identifies our 5 treatment options and 2 major diagnostic groups (early and late diagnosis). According to Tang, our treatment options included total removal without radical resection, subtotal removal without radical resection, radical resection, TS biopsy procedure followed by craniotomy, and no surgery. This grouping does not fit with our study because we had no cases of radical resection and no cases falling into the category of no surgery. Furthermore, the categories of early and late diagnosis do not apply to our study. We did not group our patients according to treatment options and time of diagnosis. During their clinical histories, patients with clival chordoma undergo several operations, often in different neurosurgical centers, and radiation therapy of various types (conventional, proton beam, brachytherapy, and so forth). Therefore, it is quite difficult to place these patients into homogeneous groups relative to treatment modality. In our introduction, we stated that although chordomas are generally regarded as slow-growing tumors, ~ 20% recur as early as 1 year after surgery, and these early recurrences can occur since the first operation. As such, the main purpose of our work was to identify the molecular features that might distinguish this subset of biologically aggressive chordomas.

Clinical Staging of the Same Patient (Sample) and use of the Chi-Square Test

Tang notes that in the studies on chordoma, CSDH, and TN, there are clinical stages in the same patient (sample). In such studies, clinical stages cannot be regarded as independent data because there is a relationship with time in terms of the data collected at each stage from the same patient. As a consequence, the researcher should not use the chi-square test when data are supposed to be independent of each other between samples. We agree with Tang and in fact did not use the chi-square test in our study of clival chordoma (see Clinical Material and Methods, Statistical Methods). Once again, Tang makes an error by specifically citing our article along with his criticisms of methods used by other researchers.

Indication of Statistical Method

Tang acknowledges that we clearly indicated the statistical methods used throughout our article.
Neurosurgical forum

Small Number of Patients

We agree with Tang that 26 patients cannot be considered a large sample size from a mere statistical point of view. From a clinical point of view, however, clival chordomas are rare tumors that account for only 0.1–0.2% of intracranial tumors. Theoretically, without any mechanism of patient selection, somewhere between 13,000 and 26,000 craniotomies for tumors must be performed to collect 26 skull base chordomas. At our center, where there is selection for skull base tumors, the series of clival chordomas presently includes 46 cases, figuring an incidence of 1.3% of all intracranial tumors, an ~10-fold higher rate than in the general population. With these numbers in mind, Tang’s statement, “For more conclusive results, the study should be extended to a much larger patient population before accepting a therapy as the treatment of choice” appears obvious and of little value.

Post Scriptum. In responding to his page proofs, Dr. Tang made several substantial changes to his original manuscript, which may influence the meaning of our response. We do believe that this unusual event is causing an unpleasant situation. Under normal circumstances, only grammatical or syntactic changes are allowed at the stage of proof correction. Conversely, Dr. Tang has erased from the original manuscript some of his own mistakes that we outlined in our response to demonstrate the superficiality of his analysis. Whether Dr. Tang, or someone else, intends to criticize a given paper, he should read it with much more attention first.

RESPONSE: We appreciate the interesting comments of Dr. Tang regarding the statistical analyses in our study.

If the diagnostic period was different between recurrent and nonrecurrent groups, there would be a confounding factor. We performed computed tomography scanning after surgery and defined a recurrent case as one in which there was an increase in the hematoma volume and a change in the hematoma density on scans within 3 months. We think such a case has some relation with surgery. We must exclude patients from the recurrence group whose computed tomography scans showed changes after 3 months, although such a case was not observed in our study.

Among the 105 patients in the study, there were only 11 documented recurrences. Based on these facts, the sample size may be small if we want to detect a small effect. That is to say, this study will only detect a large effect due to relatively small samples and events. Due to this circumstance, some of the confidence intervals became too wide, which means very unstable information. It was practically said that the number of patients should be 10–20 times the number of explanatory variables. If it is applied, the sample size of this study should be 90–180 patients. Given that the sample size of 105 is small, the sample size is considered to be in an acceptable range.

Our study was a retrospective cohort. The sample size was not considered before the study was initiated. Of course, as a post-hoc analysis, we can calculate a power to detect the effect in terms of the odds ratio. If an average proportion was 20%, the odds ratio to detect would be ≥ 5.

We performed the statistical tests in Tables 1–3 from an exploratory viewpoint. Accordingly, there seems to be no big difference in using the chi-square or Fisher exact test. We know that the Fisher exact test should be used when the expected frequency is < 5, which happens in several places from Tables 1–3. We recognize the confirmatory analysis is the finding that appeared in Table 5.

RESPONSE: Although we applaud Dr. Tang’s stated goal of improving the use of statistics in the neurosurgical literature, we read his letter with disappointment. First, we would like to point out that in 2003 Dr. Tang wrote a Letter to the Editor regarding the analyses of our paper on TN. We addressed the issues raised in that letter in a response. Dr. Tang chose to revisit the same issues in his current letter, and we will briefly address his concerns again (more detailed responses can be found in our previous letter).

The mean ± standard deviation and the median age were presented in Table 2 of TN paper. The purpose of including both the mean and median in the table is to show the distribution of the variable. Although researchers commonly use the mean ± standard deviation to present normally distributed data and the median with an interquartile range to present nonnormally distributed data, we do not believe that presenting both the mean and median violates any “statistical rule.” Additionally, it is often important to include those statistics that have been used historically to better enable the reader to compare a new study to a previous one.

According to Tang, “In Table 3 in the report on TN, if the data had been plotted on a graph and shown to be a nonnormal distribution, the graph would likely have been positively skewed.” The variables in Table 3 were all dichotomous variables and graphing them would not have contributed to the manuscript. Given that the variables were dichotomous, we did not believe it was necessary to directly state that chi-square tests were used for the comparisons. As Dr. Tang pointed out, the use of statistics in the neurosurgical literature can be confusing, and in the future clearly defining the statistics may help to educate the reader.

Tables 3–5 were based on all procedures rather than data from the first MVD surgery. We have already addressed this thoroughly in our previous response.

As explained in our response to Dr. Tang’s 2003 letter, we referred to components of the patient clinical history that were significantly associated with positive or negative outcomes as positive or negative predictors. The logistic regression was conducted with a binary outcome (complete/partial cure vs failure).

According to Tang, “The data in Table 5 consisted of ordinals. Hence, the chi-square test was no longer applicable. Instead, a Mann–Whitney test, a nonparametric method,
was used.” Although we do not disagree that a Mann–Whitney test can be used for the comparison of ordinal data, we believe that the chi-square test is still appropriate. The goal of the comparison was to test the homogeneity of the outcome in the TN diagnosis groups, and a chi-square test serves such a purpose.

In an attempt to better educate the reader we would like to explore some of what we found to be confusing and misleading statistical issues and concepts in Dr. Tang’s letter. For example, we find the assumption for the Cox regression presented by Dr. Tang questionable (“For instance, in a Cox regression model, which is most commonly used for survival analysis, the statistical assumption is a reduction in the proportional hazard.”). The Cox proportional hazard model is a semiparametric method for survival analyses. Although no assumption is made for the underlying hazard function, the model requires that the hazard ratio for any 2 samples be a constant across time periods. Thus, the basic assumption for a Cox model should be that of proportional hazards.

We are also puzzled by Dr. Tang’s use of the TN and CSDH studies as examples of the “crossover trial.” A crossover trial involves 2 or more interventions. Each participant receives some or all of the treatments in a fixed or random order. The major advantage of a crossover study is to eliminate the participant variation as each participant serves as his/her own control. In the TN and CSDH studies, there are no different treatments; only the conditions before and after surgery are considered. We think it is inappropriate to use these 2 studies as examples of a crossover trial.

Dr. Tang includes a lengthy discussion of end point; however, we found this section confusing and difficult to follow. Surprisingly, Wikipedia was cited in the discussion; however, it is not a peer-reviewed source, but a website in which anybody can change the content. Middle school students are taught not to cite Wikipedia in scholarly papers. We in the professional field should promote the appropriate use of references.

Dr. Tang discusses the use of the chi-square and Fisher exact tests and states that “there is a large disparity problem in these sample sizes; therefore, it is important to use the Fisher exact test to generate the probability values pertaining to any chi-square statistic that is obtained.” Furthermore, “If there were really 2 independent samples, each with 11 cases for a comparison, the chi-square test would be appropriate.” The use of the Fisher exact or chi-square test does not depend on the sample sizes in each group. Instead, it depends on the expected counts in each cell of an $n \times m$ table. If the expected counts are $< 5$, a Fisher exact test (for a $2 \times 2$ table) or an extended Fisher exact test (for a $n \times m$ table) should be considered. The Fisher exact test calculates the exact probability based on a hypergeometric distribution, whereas the chi-square test uses a large sample approximation to estimate the probability.

We are troubled by the section “Significance of a Relationship Between Variables Depending on Sample Size” and its discussion of the use of the ROC curve. An ROC curve is typically used to evaluate a diagnostic test, and it is obtained by plotting sensitivity against $1 - specificity$. The area under the ROC curve serves as a global assessment of the performance of the diagnostic test. We do not understand Dr. Tang’s educational point regarding sample size and the use of an ROC curve.

Post Scriptum. Dr. Tang made some significant changes to his page proofs. Our responses reflect our thoughts to his original manuscript. (DOI: 10.3171/JNS/2008/108/6/1256)

YUE-FANG CHANG, PH.D.
ELIZABETH C. TYLER-KABARA, M.D., PH.D.
MICHAEL H. HOROWITZ, M.D.
AMIN B. KASSAM, M.D.
University of Pittsburgh
Pittsburgh, Pennsylvania

Reference