Otolaryngology–Head and Neck Surgery (2007) 136, 14-18

INVITED ARTICLE

Practical guides to understanding sample size and minimal clinically important difference (MCID)

J. Gail Neely, MD, Ron J. Karni, MD, Samuel H. Engel, MD, MPH, Patrick L. Fraley, MD, Brian Nussenbaum, MD, and **Randal C. Paniello, MD,** St. Louis, MO

THE PROBLEM

One important use of the literature by a practicing otolaryngologist is to answer a question. Consequently, our literature has evolved to fulfill this need by attempting to provide more statistical analysis and higher levels of evidence; and, indeed, more level I and level II evidence is being published.[1] However, a recent study of abstracts submitted to the Academy of Otolaryngology–Head and Neck Surgery annual meeting found that while the presence of statistical analysis was associated with publication success, the level of evidence was not.[2]

Sample size is a key element in determining the usefulness of the data presented in the literature. The usefulness of a study is partially determined by sample size. Without the proper sample size, a meaningful answer to a clinical question cannot be determined. Often articles with seemingly high levels of evidence, which currently is defined only by the assembly of subjects and the allocation of interventions, do not provide definitive answers to a question because of sample size is inadequate. A recent study in the plastic surgery literature found that only 12.8% of randomized controlled trials published in the plastic surgery literature contained sample size calculations.[3] The otolaryngology literature has a lower mean sample size than other surgical subspecialty journals.[4] This leaves us with abundant literature incapable of answering the questions that we have.

The purpose of this article is to explain sample size and MCID and to illustrate their utility in the rapid critical analysis of the literature.

EXPLAINING THE SOLUTION

Sample size calculation and establishment of the minimal clinically important difference (MCID) are among the first steps in evaluating or planning a study. Sample size is defined as the number of subjects required in each arm of a study to detect a specified difference. However, sample size calculations seem to be missing from the methods and materials sections of many articles. Why is this?

Using a simple randomized clinical trial (RCT) model, Figure 1 illustrates how sample size works and why it might be unfamiliar to practicing physicians. In an RCT, estimates are made of how many potentially eligible subjects are available. The first filter is screening the potential subjects by inclusion and exclusion criteria. The resulting truly eligible subjects are then asked to participate and their consent for the trial is obtained; a portion of eligible subjects chooses not to participate. Those agreeing to the study are enrolled and randomized into smaller groups in each arm. Some of these drop out before the endpoint (the individual subject stopping time and outcome measure chosen as the primary measure for comparison between arms). A rule of thumb suggests that only 10% of potentially eligible subjects actually fulfill study criteria and complete the study.

On the other hand, in practice we treat many, or most, of the potentially eligible subjects with the same or varied treatments with no regard as to how many patients we treat or how varied they might be. Various outcomes, which may or may not be measured quantitatively, are not systematically compared. Thus, sample size does not come up for discussion in practice (Fig 1).

From the Department of Otolaryngology–Head and Neck Surgery, Washington University School of Medicine.

Dr Nussenbaum is on the Scientific Advisory Panel for Stryker Biotech.

Reprint requests: J. Gail Neely, MD, Department of Otolaryngology–Head and Neck Surgery, Washington University School of Medicine, 660 S. Euclid Avenue, Box 8115, St. Louis, MO 63110.

E-mail address: jgneely@aol.com, neelyg@ent.wustl.com.

doi:10.1016/j.otohns.2006.11.001

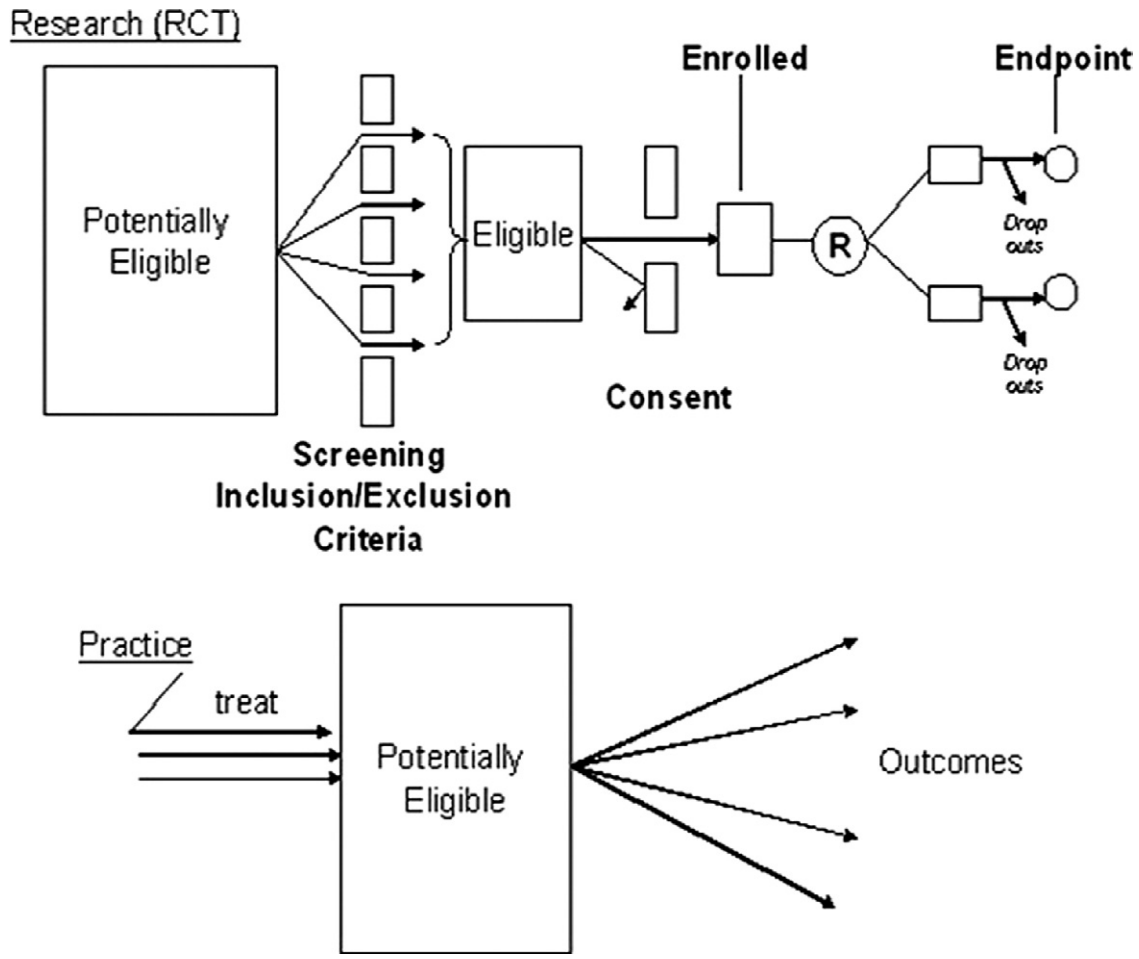

Figure 1 Illustration of the difference between clinical practice and clinical research using a randomized clinical trial (RCT) model to show how numbers of subjects markedly reduces as the presumed potentially eligible subjects for a study become filtered by inclusion/exclusion criteria and patient behavior.

In analyzing an article, in addition to determining that the study was not conducted in a biased way, it is fundamentally important to look at the numbers of subjects used if we are to feel comfortable with the authors' conclusions about comparisons of contrast or association. We need to have some assurance that these data are not falsely positive or negative, having missed a true positive result. We hope this will become apparent as one understands what sample size is all about.

EXAMPLE

What is the best way to do a tonsillectomy? The choices are nicely summarized on the website of the American Academy of Otolaryngology–Head and Neck Surgery, http://www.entnet.org/kidsent. They are: 1) cold knife (steel) dissection, 2) electrocauter, 3) harmonic scalpel, 4) radiofrequency (monopolar) ablation, 5) carbon dioxide laser, 6) microdebrider, 7) bipolar radiofrequency ablation (coblation). The contentions in the debate as to which is best seem to pivot on: postoperative bleeding, thermal injury to surrounding tissue, pain during the postoperative period, precision of cutting, time to return to school or work, postoperative sleep disturbance, need for medication, tissue healing, delayed complications, and postoperative care.

The literature on this clinical question is daunting. So how do we start? An efficient and effective way to approach the literature is to think about it first, before beginning the search. The first question is: What is meaningful to us? The next series of questions relates to how best to measure the outcome of our interest, how many subjects are needed to answer the question (sample size), and how much difference between groups we need to see to convince us to continue or change our practice.

For example, let's say we care most about postoperative pain on the day of surgery. We also know that postoperative pain can be measured on a visual analog line scale ranging from 0 to 10 cm or by 10 smiley/sad faces. So how much difference between surgical techniques would we be comfortable with? Suppose we decided that one and one half points (1.5 points) would be important to us and our patients; less of a difference might not get our attention.

We are now ready to briefly scan the literature. We want a randomized clinical trial comparing two surgical tech-

Table 1
Sample size calculations (SigmaStat 3.1)

t test	Expected difference between means
	Expected standard deviation
	Desired power (usually 0.80)
	Alpha (usually 0.05)
Paired *t* test	Change to be detected
	Expected standard deviation of change
	Desired power (usually 0.80)
	Alpha (usually 0.05)
Dichotomous proportions	Expected Group A proportion
	Expected Group B proportion
	Desired power (usually 0.80)
	Alpha (usually 0.05)
χ^2 (for more than dichotomous data)	Estimated number of observations in each cell of contingency table
	Desired power (usually 0.80)
	Alpha (usually 0.05)
ANOVA	Minimum detectable difference in means
	Expected standard deviation of residuals
	Number of groups
	Desired power (usually 0.80)
	Alpha (usually 0.05)
Correlation	Expected correlation coefficient
	Desired power (usually 0.80)
	Alpha (usually 0.05)

niques of interest that used a 10-point visual analog scale to assess the primary outcome of postoperative pain on the day of surgery. We only need one such article to give us a few facts on which to think more, before we get too involved with the extensive literature.

Noordzij and Affleck in a Triological Society candidate's thesis compared coblation with unipolar electrocautery.[5] They studied 48 subjects, randomly selecting the side coblation was to be used on. To the authors' credit, they did mention that the sample size was calculated before the study was done; however, they did not say how. Subjects were blinded as to which side received which technique and were asked to record their pain each day for each side independently for 14 days. The authors' paper found that on the day of surgery, the cautery side average pain was 5.02 and the coblation side was 3.49 (with the average standard deviation of 2.39); this was a statistically significant difference of 1.53, exceeding the threshold of 1.5 we decided was important to us. However, suppose we are also interested in the first postoperative day. Here they found a nonsignificant difference between sides of 1.04 (SD 2.56). How many subjects would need to be studied to achieve statistical significance with this difference between sides? We can turn to a statistical computer program (SigmaStat) for help. We find that 97 subjects would be required (SigmaStat calculation of sample size for *t* test: Expected difference between means 1.04; expected standard deviation 2.56; power 0.80; alpha 0.05 gives sample size [in each of two arms] of 97). What if we were interested in the seventh postoperative day? They found a nonsignificant difference of 0.57 (SD 2.5). In this situation, it would take a sample size of 303 subjects to achieve statistical significance. But note, even if 97 subjects or 303 subjects were studied to achieve a statistically significance difference between techniques, neither would pass our threshold of clinical importance.

Thus, by thinking about the problem before we looked at the literature and by retrieving one article for parameter information, we have saved ourselves a great deal of time and confusion in putting these debates into perspective. Certainly, there is a great deal more to the issue; but what we have seen from this exercise is: 1) there is an important difference between what is clinically meaningful to us (known as the minimal clinically important difference, or MCID) and what is statistically significant; 2) if a large enough sample is used, even small differences can become statistically significant; and 3) sample size and MCID are crucially important to be considered in order to put information into perspective and in order to save us precious time.

SAMPLE SIZE CALCULATION

Table 1 shows the components required to calculate sample sizes. Without getting into statistical details too deeply, let's look at what these components fundamentally represent.

First, there is an estimate of the difference expected between groups (eg, means, proportions).

Second, there is an estimate of the expected spread of values about the central tendency of the data (eg, standard deviation).

Third and fourth, a decision is made as to the limit of just how secure we wish to be in avoiding a falsely positive

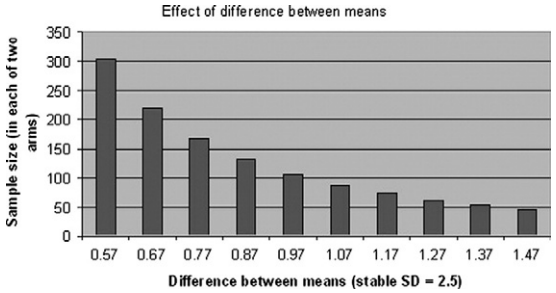

Figure 2 Illustration of how the difference between the means of two groups affects the required sample size when all other factors remain the same (*t* test of significance, power 0.80, alpha 0.05, standard deviation 2.5 on a continuous variable scale of pain ranging from 0-10). Note as the difference between the means increases, the sample size decreases.

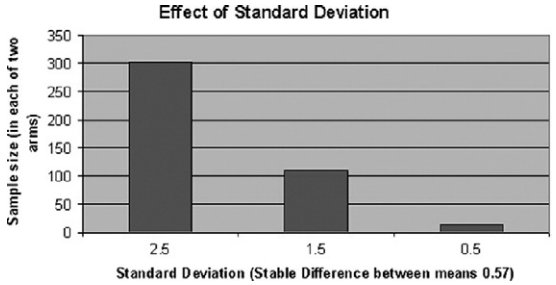

Figure 3 Illustration of how the standard deviation affects the required sample size when all other factors remain the same (*t* test of significance, power 0.80, alpha 0.05, difference between the means of 0.57). Note as the standard deviation decreases, the sample size decreases.

result (alpha) or falsely negative result (power). When we set the alpha level at 0.05, we mean that we are willing to accept a false-positive result 5% of the time. When we set the power at 0.80, we mean that we expect our data are powerful enough to avoid a false negative 80% of the time. We could set alpha at 0.01 and power at 0.90 to reduce the chance of errors even more. However, this will be costly in terms of increasing sample size.

FACTORS INFLUENCING SAMPLE SIZE

Several factors influence sample size requirements. Some of these factors are subject to change by the investigator; others are out of the control of the investigator and are determined by the biology in question.

The Difference Between Groups

The larger the difference between the central tendency of the data of each group (e.g., mean or median), the fewer subjects are needed. The biology determines this. However, the choice of a primary outcome measure is very important. For example, in the case of two equally plausible outcome measures, one might show a striking difference between groups and the other might show a much smaller difference between the study groups. In such cases, it is prudent to choose the measure showing the larger difference, unless this is irrelevant or inappropriate (Fig 2).

The Spread of Values (e.g., Standard Deviation) about the Central Tendency of the Data

If values remain close to the central tendency of the data, the smaller can be the sample. Biology determines this. However, just as with the difference between groups mentioned above, if one outcome measure is more tightly configured, it might be a better measure to select (Fig 3).

Alpha and Power Levels

If alpha is set smaller or power is desired to be larger, the larger is the required sample size. It is common practice in clinical research to set alpha at 0.05 and in bench research at 0.01. This is not absolute; however, this convention is so entrenched that changes may be hard to justify. Power, on the other hand, is often pushed to 0.90 (Fig 4).

The Scale Used for the Outcome Measure

Nominal (eg, California, Missouri, Texas, Oklahoma) and dichotomous scales (eg, man/woman, high/low) require the

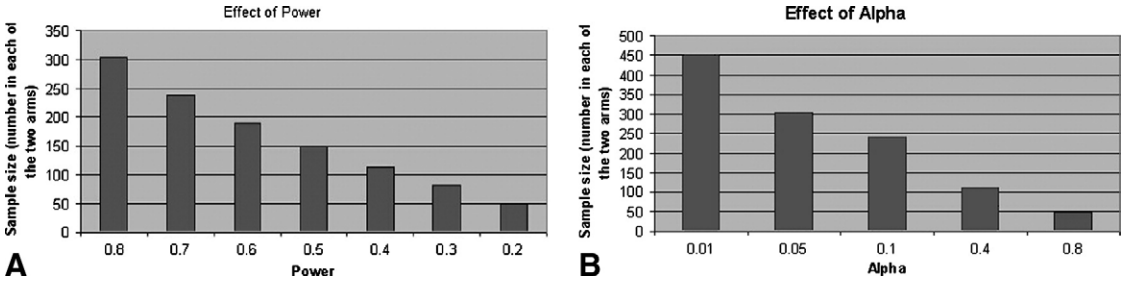

Figure 4 **(A)** Illustration of how the power affects the required sample size when all other factors remain the same (*t* test of significance, alpha 0.05, difference between the means of 0.57, standard deviation 2.5). Note as the power decreases, the sample size decreases. **(B)** Illustration of how the alpha affects the required sample size when all other factors remain the same (*t* test of significance, power 0.80, difference between the means of 0.57, standard deviation 2.5). Note as the alpha increases, the sample size decreases. The only alpha ranges generally ever considered range between 0.01 and 0.1, with 0.05 standard for clinical research.

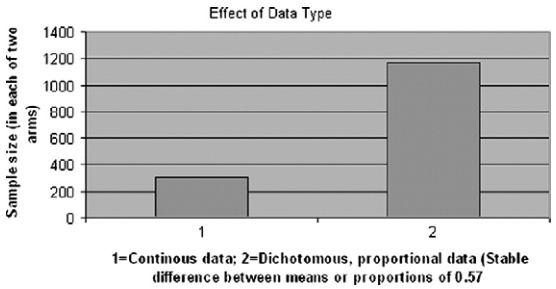

Figure 5 Illustration of how the data scale used for the outcome measure affects the required sample size. Note how much smaller the sample size is when using a continuous variable to measure the outcome than when using a dichotomous variable when keeping the difference between means (continuous variable) and the difference between proportions (dichotomous variable) the same. This generally holds true; however, there are exceptions, dependent upon the biology of the situation.

largest samples; ordinal scales (eg, Stage I-IV, Grade I-VI) require fewer subjects; and continuous scales (age, weight, serum potassium) require the fewest subjects. When it is possible to shift from a dichotomous scale to a continuous one, sample size usually decreases (Fig 5).

MINIMAL CLINICALLY IMPORTANT DIFFERENCE

Determining what is clinically meaningful is quite subjective. There are numerous articles that report multiple formal methods for calculating the minimal clinically important difference.[6] When designing a project, considerable attention should be spent on this issue because it materially affects just how many subjects are needed to achieve not only statistical significance, but clinical importance. Unfortunately, this point is rarely considered in our literature.

However, when reading the literature, we can focus primarily on what seems meaningful to us. One way is to think about it while we are making rounds, operating, and in the clinic, and by listening to our colleagues, nurses, and patients for their input on the subject. This actually can be very instructive. More formally, if one so desires, one can gather some data of one's own. For example, on the issue of postoperative pain after tonsillectomy, one could have patients record their pain on a visual analog pain scale for 1 to 2 weeks after tonsillectomy. This would give you your own data on which to calculate the mean and standard deviation. Using this data, more thinking and discussions might give a more refined answer as to what minimal difference between what you are doing and something else would cause you to seriously change what you are doing.

Discrepancy Between Calculated Sample Size and Significant Study Results

What happens when a calculated sample size is much larger or smaller than the actual study sample giving statistically significant results? Sample sizes are estimates of the minimal number of subjects required in each arm and are derived from either pilot data or the literature. If the actual difference between the study groups is larger than predicted or the spread of the data is more tightly configured about the mean, the number of subjects required to reach statistically significant results might be smaller than initially calculated. This, of course, is the dream of every investigator. It is also possible that the needed sample size might actually be larger than calculated, a nightmare of every investigator. In this situation, the actual difference between groups might prove to be smaller than predicted or the spread of the data might be more dispersed. It is because of this that interval analyses are conducted during trials, so that more or fewer subjects might be accommodated, or for ethical reasons the study must be stopped.

As the saying goes, "if it seems too good to be true, it probably is." If a statistically significant result is achieved in a trial with a sample much smaller than your experience or the previous literature would suggest, you must rule out a serious random or systematic bias in the study before accepting the article as valid.

CONCLUSION

In an effort to resolve controversy, sample size and minimal clinically important difference calculations are crucially important to consider when attempting to make sense of the literature. Unfortunately, these are not commonly considered in our literature. The purpose of this article is to raise the level of awareness on this issue, to demonstrate the utility of these calculations for the efficient and effective analysis of the literature, and to present a few facts about them for better understanding.

REFERENCES

1. Wasserman JM, Wynn R, Bash T, et al. Levels of evidence in otolaryngology journals. Otolaryngol Head Neck Surg 2006;134: 717–23.
2. Peng P, Wasserman J, Rosenfeld R. Factors influencing publication of abstracts presented at the AAO-HNS Annual Meeting. Otolaryngol Head Neck Surg 2006;135:197–203.
3. Karri V. Randomised clinical trials in plastic surgery: survey of output and quality of reporting. J Plast Reconstr Aesthet Surg 2006;59:787–96.
4. Bhattacharyya N. Peer review: studying the major otolaryngology journals. Laryngoscope 1999;109:640–4.
5. Noordzij J, Affleck B. Coblation versus unipolar electrocautery tonsillectomy: a prospective, randomized, single-blind study in adult patients. Laryngoscope 2006;116:1303–9.
6. Norman G, Sloan J, Wyrwich K. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. Med Care 2003;41:582–92.