J Oral Maxillofac Surg
78:1880-1882, 2020

# Power and Sample Size: An Opportunity to Optimize Randomized Controlled Trials in Oral and Maxillofacial Surgery Research

*Tim T. Wang, BA,* * *and Sung-Kiang Chuang, DMD, MD, DMSc*†

Randomized controlled trials (RCTs) are the gold standard for testing hypotheses in medical research. Although more resource intensive than other types of studies, RCTs generate the highest level of scientific evidence that minimizes biases. However, researchers have recently identified the need to improve the statistical rigor of RCTs in oral and maxillofacial surgery (OMS).[1] Similarly to such a trend in other surgical specialties, recent studies in OMS support the shift away from only reporting *P* values to also including 95% confidence intervals to indicate statistical significance.[1,2] However, although 95% confidence intervals help convey more information to readers to avoid a false-positive finding, they do not address the concern over a false-negative finding. The latter stems from underpowered studies, which are unfortunately common in surgical RCTs.[3] In fact, only a third of OMS RCTs even described power calculations at all.[4] As such, this article aims to highlight a critical yet underdiscussed component of OMS research: statistical power. The following provides a brief overview of statistical errors and discusses the importance of a priori sample size calculations, which can optimize resources and strengthen the quality of evidence from RCTs in OMS.

## Alpha, Beta, and Types of Errors

Scientific hypotheses are conventionally tested in the form of a null hypothesis, which the results subsequently reject or fail to reject. For example, in an RCT studying the efficacy of a new analgesic for postoperative pain control, the null hypothesis would be that the efficacy of the analgesic is the same as that of a control drug that is the current standard treatment. A study result that correctly rejects the null hypothesis when there is truly a difference in pain control efficacy between the 2 drugs constitutes a true-positive result (Table 1). Results that correctly fail to reject the null hypothesis when the 2 drugs have no difference would be true-negative results.

However, there is also the possibility of erroneous conclusions drawn from the RCT results. A false-positive result, in which the null hypothesis is rejected when there is no true difference, is defined as a type I error (alpha). A false-negative result, in which the null hypothesis fails to be rejected when there is a true difference, is a type II error (beta). Alpha and beta have historically been set at 0.05 and 0.2, respectively, which ensures that studies have a less than 5% chance of a type I error and a less than 20% chance of a type II error. This accepted alpha level is why $P < .05$ is considered statistically significant, as there is a less than 5% chance of incorrectly rejecting the null hypothesis and concluding that there is a difference when none exists. Similarly, the power of a study, which is equal to 1 – beta, indicates the probability of avoiding a type II error. Ideally, power should be 0.8 or higher, although this is not always achieved. In

*DMD Candidate, School of Dental Medicine, MPH Candidate, Perelman School of Medicine, and Associate Fellow, Leonard Davis Institute, University of Pennsylvania, Philadelphia, PA.

†Clinical Professor of Oral and Maxillofacial Surgery, School of Dental Medicine, University of Pennsylvania, Philadelphia, PA; Private Practice, Brockton Oral and Maxillofacial Surgery, Brockton, MA; and Attending, Department of Oral and Maxillofacial Surgery, Good Samaritan Medical Center, Brockton, MA.

Conflict of Interest Disclosures: None of the authors have any relevant financial relationship(s) with a commercial interest.

Address correspondence and reprint requests to Dr Chuang: PO Box 67376, Chestnut Hill Station, Chestnut Hill, Massachusetts, MA 02467; e-mail: sungkiangchuang@gmail.com

Received June 5 2020
Accepted June 6 2020



0278-2391/20/30662-5
https://doi.org/10.1016/j.joms.2020.06.020

**Table 1. ALPHA, BETA, AND STATISTICAL ERRORS**

| | Reality | |
|---|---|---|
| Test Result | + | – |
| + | True positive | Type I error |
| | | False positive |
| | | Alpha |
| – | Type II error | True negative |
| | False negative | |
| | Beta | |

*Wang and Chuang. Power and Sample Size. J Oral Maxillofac Surg 2020.*

fact, a study evaluating negative surgical RCTs from 1999 and 2009 found a third of studies to be underpowered.[3]

## Why Are Underpowered Studies Bad?

Underpowered RCTs put study conclusions at risk of a type II error. Researchers may fail to reject the null hypothesis and conclude a false-negative result. Consequently, the study will have wasted human and medical resources without generating clinically useful evidence. Even worse, if the error is not recognized, conclusions drawn from the false-negative results can have deleterious effects on patients in the clinic. Returning to the aforementioned analgesic example, incorrectly concluding that the new drug is no different than the control option when in fact it is, in reality, superior, deprives patients of the best available treatment. If we modify the analgesic example to a life-saving chemotherapy drug, the ramifications escalate to literally matters of life and death.

## Why Is Sample Size Important?

The power of a study depends on 3 factors: alpha level, magnitude of the expected difference, and sample size. Because the expected difference cannot be controlled by researchers and alpha is conventionally maintained at 0.05, the sample size becomes the crucial element to ensure studies are appropriately powered. To this end, sample size calculations before RCTs are critical. An appropriately sized sample will not waste extra resources but will enable the researchers to make proper conclusions from their results.

## Which Factors Influence Sample Size Needed?

Several factors influence the sample size calculation.[5] In terms of RCT design, the ratio at which individuals are allocated into the treatment arms and the type of outcome variable measured—either binary or continuous—are 2 major factors. In addition, the expected effect size is inversely correlated with the sample size needed. Finally, larger variability within the sample observations necessitates a larger sample size.

To determine these components, there are several strategies that researchers can use. For example, researchers can pilot the experiment with a small sample to determine expected variability and effect size. They also can conduct a systematic review and meta-analysis to synthesize other published evidence on analogous data. Finally, they can reference expert experience on the topic. Ideally, researchers will use some or all of these approaches when calculating the sample size. In addition, they should factor in the possibility of participant dropout and noncompliance during the course of the experiment. Farrokhyar et al[5] have provided additional information on the equation and rationale for calculating power.

## Multicenter RCTs

Once a sample size estimate is calculated, it is important to reach this threshold or else an RCT risks being underpowered. If circumstances make it impossible to reach the target sample size at a single institution, researchers can consider multi-institution RCTs. Although the same statistical principles apply, RCTs at multiple centers have both pros and cons. Involving multiple institutions increases the heterogeneity of the sample. This can increase the external validity, or generalizability, of the results to a wider population. However, the study's internal validity may be decreased because the experimental design may not be as well controlled; it may be practically impossible to maintain identical conditions across different institutions. These tradeoffs in experimental design should be carefully considered and reported in the final study.

This article highlights the importance of calculating the sample size to appropriately power RCTs. Conducting and reporting these calculations in the final article should be encouraged because doing so can strengthen the statistical validity of the research. This article by no means is an adequate resource to guide a full-scale RCT, and researchers without prior experience should consult experts in biostatistics before the trial and throughout its course. Researchers also can participate in the American Association of Oral and Maxillofacial Surgeons workshop on research methods. The recent attention on statistical rigor is encouraging and should be built on to further improve the quality of OMS research. In the long-term, OMS should expand initiatives to conduct large-scale RCTs to generate and

subsequently implement top-quality evidence to improve surgical care.

## References


1. Susarla SM, Dodson TB, Cheng KL: Do academic oral and maxillofacial surgeons comply with best practices for reporting the results of randomized clinical trials? J Oral Maxillofac Surg 78:771, 2020
2. Karadaghy OA, Hong H, Scott-Wittenborn N, et al: Reporting of effect size and confidence intervals in *JAMA Otolaryngology-Head & Neck Surgery*. JAMA Otolaryngol Head Neck Surg 143:1075, 2017
3. Ali UA, ten Hove JR, Reiber BM, et al: Sample size of surgical randomized controlled trials: A lack of improvement over time. J Surg Res 228:1, 2018
4. Kyzas PA: Evidence-based oral and maxillofacial surgery. J Oral Maxillofacial Surg 66:973, 2008
5. Farrokhyar F, Reddy D, Poolman R, Bhandari M: Why perform a priori sample size calculation? Can J Surg 56:207, 2013