Sample size estimation: An overview with applications to orthodontic clinical trial designs

Nikolaos Pandis,a Argy Polychronopoulou,b and Theodore Eliadesc
Corfu and Athens, Greece, and Zurich, Switzerland

Proper sample size estimation is an important part of clinical trial methodology and closely related to the precision and power of the trial’s results. Trials with sufficient sample sizes are scientifically and ethically justified and more credible compared with trials with insufficient sizes. Planning clinical trials with inadequate sample sizes might be considered as a waste of time and resources, as well as unethical, since patients might be enrolled in a study in which the expected results will not be trusted and are unlikely to have an impact on clinical practice.

Because of the low emphasis of sample size calculation in clinical trials in orthodontics, it is the objective of this article to introduce the orthodontic clinician to the importance and the general principles of sample size calculations for randomized controlled trials to serve as guidance for study designs and as a tool for quality assessment when reviewing published clinical trials in our specialty. Examples of calculations are shown for 2-arm parallel trials applicable to orthodontics. The working examples are analyzed, and the implications of design or inherent complexities in each category are discussed. (Am J Orthod Dentofacial Orthop 2011;140:e141-e146)

Clinical trials in orthodontics aim to assess prospectively associations between treatment interventions and outcomes, and to determine which is the better intervention for the desired outcome. For example, failure rates of brackets bonded with different composites, survival of different types of lingual retainers over several years, treatment efficiency of bracket systems, effectiveness of various functional appliances, and combinations of treatments with fixed and removable appliance can be compared.

It is important that a certain methodology is followed so that the trial possesses what is called in clinical trial language “internal validity” and produces precise results. Some important aspects of internal validity or methodologic quality are randomization, blinding, and reduction of loss to follow-up.1-3 Additionally, for the results to be credible, it is necessary to have enough participants in the trial to have a good chance of detecting a clinically important treatment difference if there is one, or to conclude with reasonable confidence that there is no difference if the results do not show it. Often, trials fail to show an effect of a particular intervention, even when an effect exists, because of the small size of the trial; this leads to low power, and, as it has been argued, “absence of evidence is no evidence of absence.”4,5

Every clinical trial should have good power, and there is a close relationship between power and sample size. Usually, as the sample size increases, study power is also expected to increase; however, a balance between power, a clinically important difference, trial feasibility, and credibility is required.

It has been reported that in medical6 and dental7-12 journals reporting for sample size calculation is suboptimal. Because of the low emphasis on sample size calculation in dental medicine, it is our objective to introduce the orthodontic clinician to the importance and the general principles of sample size calculations for clinical trials as guidance for study designs and as a tool for quality assessment when reviewing published clinical trials in our specialty. This article is by no means exhaustive. It focuses on trial designs applicable to orthodontics and presents sample size calculations with only proportions or means. Sample calculations with risk, odds, and hazard ratios are not included.

The key factors for sample size determination will be introduced, and then sample size calculations with...
examples of common trial designs used in orthodontic research will be presented.

**KEY FACTORS IN DETERMINING SAMPLE SIZE**

Before we proceed with sample calculations we must define the following: (1) the research question, (2) the principal outcome measure of the trial (ie, continuous or categorical), (3) the type of analysis and study design that will be used, (4) the expected proportion or mean value of the selected outcome in the comparison or control group, (5) the minimum clinically important difference between treatment arms that we would like to detect, (6) the variance (for continuous outcomes), and (7) the significance level (α) and the degree of certainty (β) that we want to detect for the treatment difference.

We must accept some risks of a false-negative result (type II error) and also of a false-positive result (type I error); otherwise, we would need an infinitely large study. Sample calculations involve some judgment, and some compromise is required between strict statistical criteria and practicality.

As the minimum clinically important difference to be detected decreases, the sample size increases and vice versa. Figure 1 shows the effect of the size of the difference to be detected between treatment arms on the required sample size and power using simulated data.

As the variance (for continuous outcomes) increases, the required sample size increases. Figure 2 shows the effect of the variance on the required sample size and power using simulated data.

As the number of events (outcomes) become rare, the required sample increases.

As the desired power level increases (power = 1 − β) or the probability of type II error (β) and type I error (α) decreases, the required sample size increases and vice versa.

Sample calculations are performed at the design stage of the trial, and beforehand knowledge of the range of values for some calculation parameters is required. Information on those parameter values might be obtained from previous research, publications, or pilot studies. Sample calculations are meaningless after trial conclusion; at this stage, power and precision can be assessed by using the produced confidence intervals.

**SAMPLE CALCULATIONS FOR CLINICAL TRIALS: PARALLEL ARM DESIGNS**

First we will conduct a sample calculations for a trial with qualitative outcomes, a 1:1 allocation ratio, and a 2-sided test.

---

**Fig 1.** Sample size vs power for different minimum differences to be detected between treatment groups. A sample size of 100 gives a power of 80% for a minimum difference (in proportions) to be detected of 0.15, whereas, for a minimum difference of 0.05, the power is only 10%.

**Fig 2.** Sample size vs power for the same minimum difference (10 units) to be detected between treatment groups with variable standard deviations (sd) (for each line, the standard deviation is assumed to be the same for both treatment groups). A sample size of 65 gives a power of 80% for a minimum mean difference to be detected of 10 units and sd1 = sd2 = 20 units, whereas for a minimum mean difference of 10 units and sd1 = sd2 = 40, the power is only about 30%.

---

We are interested in conducting a trial to compare failures of lingual retainers bonded with chemically cured or a light-cured adhesive, and to determine whether there is a difference at the 0.05 statistical level with 90% power. The procedure for sample calculation for such a study will be as follows.
1. We must decide what is an acceptable difference to be observed that is of clinical importance. Selecting a difference that is too small and not clinically important will increase the required sample size to impractical levels; at the same time, the trial will give results that, although significant, might have no clinical importance. Let us assume that we consider a 10% difference in the proportion of failures to be an important clinical difference.

2. We must make assumptions regarding the expected proportion of failures in the control arm, the chemically cured adhesive in this example. Sources that will help us to determine the expected proportion of failures in the control arm could be published studies or a pilot of the trial we are designing. Let us assume that we have found a study that reported on the proportion of failures for the chemically cured adhesive we are planning to use and that this proportion of failures is 20%.

So far, we have the proportion of failures in the control arm (20%), the minimum difference (10%) to be detected, and the desired significance and power levels (5% and 90%). To carry out this calculation, we will use the formula described by Pocock13 that assumes independently distributed outcomes, equal numbers of participants per arm, no losses to follow-up, and no continuity corrections:

\[ n = f(\alpha, \beta) \times \frac{\pi_1 \times (1 - \pi_1) + \pi_2 \times (1 - \pi_2)}{(\pi_1 - \pi_2)^2} \]

Formula 1 is for sample calculation for proportions, where \( n \) is the required sample size per trial arm, \( \pi_1 \) is the expected percentage of failures in those on standard treatment, \( \pi_2 \) is the expected percentage of failures in those on the new treatment, and \( f(\alpha, \beta) \) is a function of \( \alpha \) and \( \beta \) derived from the standard normal distribution, with its values given in the Table.13

By substituting the selected values in formula 1 above:

\[ n = 10.5 \times \frac{0.2 \times (1 - 0.2) + 0.1 \times (1 - 0.1)}{(0.2 - 0.1)^2} = 262.5 \]

This indicates that the required sample size would be 263 (rounded up) per treatment arm, for a total of 526 patients. If we reduce the power to 80%, the required sample size would be:

\[ n = 7.85 \times \frac{0.2 \times (1 - 0.2) + 0.1 \times (1 - 0.1)}{(0.2 - 0.1)^2} \]

\[ = 196.25, \text{ for a total of } 394. \]

By playing with different scenarios, we can see that the required number of patients increases as the proportion of the events in the 2 arms is closer to the 0.5 value, the difference to be detected decreases, the level of significance decreases, and the power increases.

Next we will perform a sample calculation for a trial with normally distributed quantitative outcomes, a 1:1 allocation ratio, and a 2-sided test.

In this example, we are interested in comparing the number of days it takes to reach dental arch alignment with a conventional appliance and a self-ligating appliance. For the sample calculations for comparison of proportions, certain assumptions and decisions should be made. We assume that, for the conventional bracket, 200 days are needed to reach alignment (\( \mu_1 \), comparison or control group), with a standard deviation (sd) of 50 days. We want to calculate the required sample to be able to detect a difference of 40 days between the 2 bracket systems at significance level \( \alpha = 0.05 \) and power \( = 1 - \beta = 90\% \). The extra parameter that must be defined in this scenario is the standard deviations of the means to be compared.

\[ n = f(\alpha, \beta) \times \frac{2sd^2}{(\mu_1 - \mu_2)^2} \]

Formula 2 is for sample calculation for 2 means according to Pocock13; \( \mu_1 \) is the expected mean number of days to reach alignment with the standard treatment, \( \mu_2 \) is the expected mean number of days to reach alignment with the alternative treatment, sd is the standard deviation of the response (assumed to be the same on both arms), \( \alpha \) is type I error (significance level), and \( \beta \) is type II error \( (1 - \beta = \text{power}) \) and \( f(\alpha, \beta) \) is a function of \( \alpha \) and \( \beta \) derived from the standard normal distribution, with its values given in the Table.

By substituting the appropriate values into formula 2, we get:

\[ n = 10.5 \times \frac{2 \times 50^2}{40^2} = 32.8, \text{ for a total of 66 (rounded up)} \]

We assume that observations are independent, the numbers of participants per trial arm are the same, and there are no losses to follow-up. By experimenting with the formula, we can see that the required sample

<table>
<thead>
<tr>
<th>( \alpha )</th>
<th>0.05</th>
<th>0.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>13.0</td>
<td>17.8</td>
</tr>
<tr>
<td>0.1</td>
<td>10.5</td>
<td>14.9</td>
</tr>
<tr>
<td>0.2</td>
<td>7.85</td>
<td>11.7</td>
</tr>
<tr>
<td>0.5</td>
<td>3.84</td>
<td>6.63</td>
</tr>
</tbody>
</table>

(95% power) (90% power) (80% power) (50% power)
size increases when the required difference to be detected is decreased, the power level is increased, the \( \alpha \) level is decreased, or the variance is increased and vice versa.

When patients are lost during follow-up, apart from the possibility of postrandomization bias, the power of the trial decreases. Therefore, it would be prudent to account for the expected losses to follow-up by appropriately increasing the initial sample size accordingly.

THE ALTMAN NOMOGRAM

Altman\(^{17}\) created an easy-to-use nomogram for quick sample calculations for continuous outcomes for equal-size 2-arm trials. In Figure 3, on the left vertical line, is the effect size (standardized difference) for which the sample size is required at the prespecified significance level (shown on the diagonal line in the center) and power level (on the right vertical axis). The effect size in terms of the standardized difference is defined as the ratio of the minimum difference to be detected to the standard deviation. For example, when using quantitative outcomes, if the difference to be detected is 40 units and the standard deviation is 50 units, then the standardized difference is 40/50 = 0.8. If we draw a line from the left vertical line starting at 0.8 and connect it with the value of 0.90 on the right vertical line representing power, the point where the connecting line crosses the diagonal sample line indicates the total required sample size for both arms. In this example, the line crosses the 0.05 line between 60 and 70; this is close to the value of 66 found with formula 2. The corresponding sample size at \( \alpha = 0.01 \) is about 90, again close to the result of using formula 2, which gives a value of 93 for both arms. This method can also be applied for paired designs in which the required sample is approximately halved.\(^{17,18}\)

We can also calculate the required sample for quantitative paired outcomes (split-mouth design), with a 1:1 allocation ratio assuming normality in a 2-sided test.

For example, let us assume that we want to evaluate anchorage loss during space closure in Class II Division I maxillary premolar extraction patients by using for anchorage either a microimplant on 1 side or just the dental unit from second premolar to second molar on the contralateral side. In this situation, we are interested to see whether the difference in anchorage loss in millimeters after complete space closure between the 2 types of anchorage is different from zero. This is a paired observation case, where the participant is serving as the control, since both intervention and comparison treatments are applied on each patient. This design is more efficient, since the sites that receive the interventions are more similar, thus reducing variance and therefore sample size requirements. The paired design can be effectively reduced to a 1-sample design where the mean and the standard deviation of the difference between the 2 treatment arms are used for the sample calculation. Thus, formula 2 can be used, with the difference that we have only the variance of the difference rather than the variance of each intervention arm, thus reducing the formula to:

\[
\begin{align*}
n &= f(\alpha, \beta) \times \frac{s_d^2}{(\mu_1 - \mu_2)^2} \\
&= 10.5 \times \frac{3^2}{3^2} = 10.5
\end{align*}
\]

Formula 3 is for sample calculation for 2 paired means. It was adapted from the study of Machin and Fayers\(^{19}\) to correspond to formula 2, where \( s_d \) is the standard deviation of the within-person differences of \( \mu_1 - \mu_2 \).

These calculations assume that the within-persons correlations are greater than the between-persons correlations.

If the minimum difference to be detected is \( \mu_1 - \mu_2 = 3 \) mm, and \( s_d = 3 \) mm at the 5% significance level with 90% power, the required sample size for this paired design would be:

\[
\begin{align*}
n &= f(\alpha, \beta) \times \frac{s_d^2}{(\mu_1 - \mu_2)^2} \\
&= 10.5 \times \frac{3^2}{3^2} = 10.5
\end{align*}
\]

Therefore, 11 per side for a total of 22 sides (rounded up) or 11 patients, since each patient contributes 2 sides.
The corresponding sample size for an unpaired design with the same parameters would be 22 patients. We can see that this split-mouth design is efficient, since it reduces the sample size to half compared with the design that will randomize the 2 interventions to separate patients.

Split-mouth designs are not appropriate for all interventions and especially when the therapy on a quadrant could affect the outcome on another quadrant, or when it is difficult to find appropriate and similar pairs of quadrants or teeth.20-26

OTHER TRIAL DESIGNS

In certain situations, it is possible to evaluate 2 or more interventions or parameters simultaneously in a single trial. This can be accomplished with a factorial design, which takes the form of $2 \times 2$ in the simplest form. The advantages of the factorial design are related to the fact that 2 or more parameters can be assessed at the same time, thus creating a more efficient trial in terms of resources, including sample size, compared with separate trials that assess each parameter. However, certain assumptions must not be violated for the factorial design to be appropriate. The assumptions are that the 2 treatments can be combined, and there is no interaction (when the effect of 1 treatment on an outcome is not modified according to the level of a second treatment).27

In certain situations, it is not possible to use individuals as the randomization unit, but we must randomize to clusters (groups) consisting of a few or many individuals or units that share some common characteristics. Clusters may be families, schools, communities, or orthodontic practices, but also orthodontic patients contributing several teeth (several dental units clustered in a patient) or multiple measurements over time on the same person.28,29

Clustered designs in sample calculation and data analysis from individually randomized trials, which assume that observations are independent. In cluster randomized trials, observations are correlated with less information obtained per individual, resulting in loss of power compared with individual randomized trials. In cluster randomized trials, an increased sample size is required to compensate for loss of information (loss of precision or power) because of within-cluster data correlation.10-12

Split-mouth designs with clustering effects might be appropriate when subjects vary considerably in their baseline characteristics and responses to treatment. In conventional designs (with each patient allocated to 1 treatment group), this variability between subjects can often require large numbers to provide reliable evidence of a treatment effect. As pointed out earlier, in split-mouth design trials, each patient acts as his or her own control, and treatment comparisons are made within subjects, allowing for a more precise estimate of treatment effects and thus requiring a smaller sample size. Additionally, a split-mouth design using patient quadrants as clusters is more efficient compared with using patients as clusters because of the reduced within-patient variability.32-37

Finally, the analysis of covariance approach (ANCOVA) design requires a smaller sample size; this is the before-and-after design when there is a high correlation between the baseline and final continuous outcome measurements on the same subject. The higher the correlation between the 2 measurements, the smaller the required sample size and vice versa.38,39

CONCLUSIONS

1. Sample calculations are an important part of solid clinical trial methodology.
2. Sample calculations should balance scientific validity with feasibility.
3. Important differences between intervention arms might go undetected only because sample size and power are low.
4. Sample calculations make certain assumptions, and information from previous studies or pilots are helpful for reliable sample calculations.
5. Trial design characteristics such as clustering and pairing should be considered during sample size calculations.
6. Clustering requires a larger sample size compared with nonclustering, whereas pairing decreases the required sample size compared with nonpairing.

REFERENCES