



## Review Paper

# A review of sample size determination in randomised controlled studies

Kazeem A. Osulale\*, Adesola Z. Musa and Oliver E. Ezechi

Nigerian Institute of Medical Research, 6 Edmund Crescent, P.M.B. 2013, Yaba, Lagos, Nigeria  
whereisqosimadewale@gmail.com

Available online at: [www.isca.in](http://www.isca.in), [www.isca.me](http://www.isca.me)

Received 20<sup>th</sup> February 2022, revised 2<sup>nd</sup> February 2023 accepted 16<sup>th</sup> April 2023

## Abstract

*This review discusses some statistical concepts relevant to understanding and estimating sample size for randomised controlled trials (RCTs). A RCT is still the best way to compare the effects of treatments. This concept is illustrated with illustrative examples of equivalence, non-inferiority and superiority RCTs. The RCTs can be designed and conducted as superiority studies. Randomised clinical trials can be used in establishing that a proposed intervention is equivalent or noninferior to standard therapy rather than superior, meaning that the trials can be conducted as such studies. This research differs significantly in several methodological aspects because they focus on different goals. To be honest, awareness of the differences in clinical trial methodologies is often limited. This paper provides an overview of the methodology of this type of research which includes planning, execution, analysis, and reporting of trial. We hope that review article of this nature will be useful to biomedical researchers and other scientists in estimating sample sizes for studies in their various disciplines.*

## Keywords:

## Introduction

Randomised clinical studies are important to evaluate alternative treatments or interventions. In many applications, two or more forms of therapy in a clinical study can be compared. One is the therapy for control group and the other is the therapy for experimental group. In fact, a randomised clinical trial (RCT) remains the best method used to compare effect of therapies<sup>1,2</sup>.

Determination of sample size is important in randomised controlled studies<sup>3</sup>. It is unethical to conduct a CT with a very small sample size and a very large sample size, which can be worthless if the study is not adequately powered to show a meaningful difference and the researcher might give participants a therapy that might have been established to be inferior<sup>4</sup>. A too small sample may fail to produce a conclusive and reliable results of a study while on the other hand, a too large sample for a study may be a waste of resources and time<sup>5,6</sup>.

A very small sample may not lead to decisive and reliable results from a study, while on the other hand too large a sample for a study may be a waste of resources and time<sup>5,6</sup>. The participants required to conduct a clinical trial vary over several orders of magnitude. Rather than selecting a random sample size for the study, the investigator should consider both the variance in treatment response and the level of presumed treatment efficacy to determine the number of participants to use in the study to answer a scientific question<sup>7</sup>.

Standard error (SE) of results reduces with the size of the sample and therefore the accuracy and power of the study is

increased<sup>8,9</sup>. Nevertheless, the investigator may have a limited time or amount or a group of patients. However, under known assumptions, a medical statistician can be of help in calculating the appropriate sample size for the study. The possibility of achieving an effect of a given magnitude should be calculated for a given number of patients. If the only outcome variable is success or failure, the statistician must calculate the number of expected successes in the groups to determine the possible differences in potential clinical association between them.

The calculation of sample size for United Kingdom Medical Research Council (UKMRC) randomised gastric surgery study comparing conventional versus radical surgery based on the consensus of the surgical members of the design team that suggested that five-year survival rate of conventional 0.20 to 0.34 for radical surgery could be realistic and medically significant<sup>26</sup>. Sample size of 400 patients was used based on these survival rates<sup>10</sup>. Different statistical formulae exist for estimating sample sizes of various kinds of parameters and variables with their respective study designs. Online sample size calculators and software are also available to calculate a sample size once researchers understand better the basic statistical concepts required for this purpose.

**Components and basic statistical concepts of sample size estimation:** The investigator is expected to understand the relevant components of the statistical concepts needed to determine the number of patients for a randomised controlled trial<sup>11-13</sup>.

**Null Hypothesis and Alternative Hypothesis:** Hypotheses are usually made in randomised controlled trials to determine if

there is a significant change between the intervention and control groups. The hypothesis ( $H_0$ -null hypothesis) states that no difference is observed and is rejected if the p-value is  $< 0.05$ , or worse and acceptable if the p-value is  $> 0.05$ . The hypothesis presented against  $H_0$  is the alternative hypothesis ( $H_A$ ), indicating the observed difference between the treatments<sup>14</sup>. The researcher must clearly define  $H_0$  with an appropriate sample size for the study. However, the test size must be set beforehand when estimating a sample size.

### Acceptable significance level

The acceptable level of significance is denoted by  $\alpha$  and it is a type-1 error rate,  $\alpha = P$  [Type-1 error]. It is the probability of rejection of a true  $H_0$ . In medical research, it is common to set significance levels of  $\alpha = 5\%$  ( $\alpha/P = 0.05$ ) or  $1\%$  ( $\alpha/P = 0.01$ ), meaning that the research admits that 5% or 1% probability of the outcome being observable is due to chance rather than the intervention. Relative confidence levels are significant for each level: 95% CI for 5% ( $\alpha/P = 0.05$ ) and 99% CI for 1% ( $\alpha/P = 0.01$ ) for significance levels.

**Statistical power of a study:** The probability of rejecting  $H_0$ -null hypothesis when in fact it is not true is termed study power. An increased statistical power of a study minimizes a likelihood of committing a Type-2 ( $\beta$ ) error and thus decreases the risk of having a false negative result<sup>15</sup>. Power of a study is therefore denoted by  $1-\beta$ . Quite a number of clinical studies consider 80% (0.8) power or more to detect a significant difference. A study powered at 80.0% implies a likelihood of 20.0% that a significant difference cannot be detected even though it is present, and this is also consistent with the 90.0% power of large studies.

**Underlying population event rate:** The expected primary event or prevalence in the experimental or control group will be determined as stated in the literature or previous studies by other means which include an observational cohort. The investigator should exercise caution when reporting event rates and it is even best to consider adjusting sample sizes in any ongoing studies, which is important in the circumstances that the overall event rate is unexpectedly low. For example, the prevalence of antiretroviral drugs in the treatment of HIV patients should be known in advance while investigating the association between antiretroviral drugs and HIV.

**Effect size:** The expected effect size in a study is the absolute difference, that is, the difference between the frequency of events in the control and intervention groups. It can also be expressed as a comparative decrease, that is, as a corresponding change in the event rate with treatment. If the event rates are 8.30% and 6.30% in the control and intervention arms, the absolute difference is 2.0% implying that a relative reduction of the intervention is 2.0%. Cohen's guideline for effect size states that an effect size  $< 10.0\%$  (0.1) is a small effect, 30.0% (0.3)–50.0% (0.5), a medium effect, and  $> 50.0\%$  (0.5) is considered a large effect. Therefore, an effect size of 0.5 is generally used without modification and represents a moderate to large

differential effect. However, an effect size is inversely proportional to sample size. A smaller number of study participants is required when the effect size is large and vice versa. In fact, effect size is an important variable in calculating a sample size. This is often referred to as expected benefit, but it is usually taken as the magnitude of the effect that makes it worth using a new treatment to replace the previous one.

**Margin of error ( $M_e$ ):** Margin of error is the random sampling error that represents the probability that the sample results deviate from the population. For example, assume a 23% prevalence of antimicrobial resistance among children in a study sample and set the error rate to 10%. This means that the range of antimicrobial resistance in the pediatric population is 13% and 33% prevalence of antimicrobial resistance in the child population.



**Root mean-square deviation of the outcome measure:** It is important to report the root mean-square deviation of the outcome measure also known as the standard deviation for measured data. This is inferred from the literature or previous studies that have used the same measure. Care should be taken to use standard deviation rather than standard error. This should be the squared deviation of the outcome measure and not the difference in outcome measure between the intervention and the control. In most cases, only a few articles provided effect estimates with a 95% CI. The standard deviation (SD) can be calculated from the confidence interval (CI). The difference between UL (upper bound) and LL (lower bound) of CI is four times the standard error (SE) and therefore SD can be expressed as the product of SE and the square root of n, i.e.  $SD = SE \cdot \sqrt{n}$ .

### One-tailed and two-tailed hypotheses

The choice of one or two-tailed test is dependent on the purpose of the study. For instance, if research shows that a new drug is more efficacious in lowering blood pressure; then a one-tailed test may be appropriate to test the hypothesis, but if you are uncertain whether radical surgery is more or less effective in improving survival than conventional surgery<sup>16</sup>, then it is more appropriate to use a two-tailed hypothesis<sup>15</sup>. The values of parameter for both tests are identical but the distinction is in the critical ratio (z-score). The two hypotheses have their respective z-score, as shown in Table-1<sup>17</sup>.

**Design effect:** Design effect (deff) is a significant component in the estimation of sample size<sup>18,19</sup>. The formula applied in the calculation of sample size helps estimate an appropriate sample size when simple random sampling (srs) method is used in research. However, if a srs technique cannot be used, the estimated sample may not be adequate, and the sample size is adjusted using a deff in such cases. The design effect is the ratio of the variance expected in clustered sampling to the variance expected in srs. The design effect is generally greater than or equal to 1. Therefore, we assume  $deff = 2.0$  in the cluster design.

**Table-1:** Standardized normal distribution table<sup>16</sup>.

Two-sided probability 		One-sided probability 			
$Z_{2\alpha}$	$2\alpha$	$Z_\alpha$ or $Z_\beta$	$\alpha$ or $\beta$	$Z_\alpha$ or $Z_\beta$	$\alpha$ or $\beta$
3.72	0.0002	3.72	0.0001	0.00	0.50
3.29	0.001	3.29	0.0005	-0.13	0.55
3.09	0.002	3.09	0.001	-0.25	0.60
2.58	0.01	2.58	0.005	-0.39	0.65
2.33	0.02	2.33	0.010	-0.52	0.70
1.96	0.05	1.96	0.025	-0.67	0.75
1.64	0.1	1.64	0.05	-0.84	0.80
1.28	0.2	1.28	0.10	-1.04	0.85
1.04	0.3	1.04	0.15	-1.28	0.90
0.84	0.4	0.84	0.20	-1.64	0.95
0.67	0.5	0.67	0.25	-1.96	0.975
0.52	0.6	0.52	0.30	-2.33	0.990
0.39	0.7	0.39	0.35	-2.58	0.995
0.25	0.8	0.25	0.40	-3.09	0.999
0.13	0.9	0.13	0.45	-3.29	0.9995
0.00	1.0	0.00	0.50	-3.72	0.9999

*Note.* The total area under the normal distribution curve is one. The area under a given part of the curve gives the probability of an observation being in that part. The  $y$ -axis indicates the "probability density", which is highest in the middle of the curve and decreases in either direction toward the tails of the curve. The normal distribution is symmetric, i.e. the probability from  $Z$  to plus infinity (right side of the table) is the same as from  $-Z$  to  $-\infty$ . The right side of the table gives the one-sided probability from a given  $Z$ -value on the  $x$ -axis to  $+\infty$ . The left side of the table gives the two-sided probability as the sum of the probability from a given positive  $Z$ -value to  $+\infty$  and the probability from the corresponding negative  $Z$ -value to  $-\infty$ .

Sometimes the goal of randomised clinical trials is to establish that a proposed intervention is equivalent or no inferior rather than superior to a standard therapy<sup>20</sup>. These studies have different goals and differ significantly in methodological aspects<sup>21</sup>. Knowledge of methodological differences in randomised clinical trials is quite limited. If no significant difference is detected, for example, between the treatments in a superiority study, it indicates that the treatments have similar or equivalent effect<sup>14,22-27</sup>. However, such a conclusion is misleading due to the possibility of missing an effect of clinical importance with a very small number of participants in the study. This review discusses various methods used to estimate sample size for this type and other related studies<sup>28</sup>.

**Superiority trials:** A key aspect in planning randomised clinical trials (RCTs) is determining the required sample size. A superiority study has a goal of proving that an intervention is superior compared to a standard therapy. To calculate sample size, the researcher must consider some important questions, such as: To what extent will an intervention be more effective than the established therapy? The additional effect of the intervention relative to the standard therapy is known as the least relevant difference (LRD) or sometimes called a clinical significance,  $\Delta$ .

The next question is, to what extent do random factors affect the difference in effect between the two groups<sup>16</sup>? As with any other biological measure, treatment effects are highly random and must be determined and accounted for. The sample SD or variance,  $S^2$  describes the degree of dispersion and the variance of the effect variables can be known either from a previously conducted or pilot studies.

The superiority trial must reveal the actual differential effect between the treatments as precisely as possible<sup>16</sup>. It is also important to indicate the large risks of Type-1( $\alpha$ ) and Type-2

( $\beta$ ) errors that will be acceptable in the analysis due to random differences that may cause discrepancies between the result of the final analysis and the actual difference that leads to erroneous results. Ideally,  $\alpha$  and  $\beta$  risks should be close to 0 (zero) but it requires extensive testing. However, limited resources and sample size make it necessary to accept a low risk of type-1 and type-2 errors. It would be of interest in many situations to determine benefits and harms of a proposed intervention relative to a control therapy, that is, "two-tailed" tests to determine the difference between an up and down association would be of interest. Therefore, we would define the risk of type-1 error as  $\alpha$  up +  $\alpha$  down, that is,  $2\alpha = 0.05$ . The risk associated with type-2 error,  $\beta$  is usually between 0.1(10%) and 0.2 (20%). The type-2 error,  $\beta$  risk tends to be one-sided following a given  $\Delta$  value that is always above or below zero ( $H_0$ ). A smaller  $\beta$  gives a higher probability complement  $1 - \beta$  of not rejecting  $H_\Delta$  when it is indeed correct. The sample size required for the intervention and control groups can be calculated with known values of  $\Delta$ ,  $S^2$ ,  $\alpha$  and  $\beta$ , using a relatively simple general formula<sup>13</sup> defined below:

$$n_0 = \frac{(z_{2\alpha} + z_\beta)^2 \cdot S^2}{\Delta^2} \tag{1}$$

Where:  $S^2$  is the variance =  $p_1(1 - p_1) + p_2(1 - p_2)$  and  $n_0$  is sample size to be calculated;  $Z_{2\alpha}$  is a statistic that defines the required confidence level; 1.96,  $z_\beta$  is the desired power (80%),  $p_1$  is the estimated prevalence at baseline,  $p_2$  is the estimated response rate of the new intervention,  $\Delta$  is taken as the difference between the expected prevalence at baseline and the response rate of the new intervention.

The right side of Table-1 will be applied if the investigator wishes to determine a difference in one direction and  $z_{2\alpha}$  would be replaced with  $z_\alpha$  in Equation (1). The formula in Equation (1) gives a good estimation of the sample size required for the

study. A study with two groups of equal size has its overall sample size to be  $2n$ . Examples in this section were drawn from an existing work found useful in this review<sup>17</sup>.

**Case 1<sup>16</sup>:** In naive cases of chronic hepatitis C, pegylated genotype 1 interferon plus ribavirin for three months induces a sustained virological response in approximately 40.0%<sup>16</sup>. The investigator wishes to verify whether a new therapeutic regimen can increase the sustained response in this type of patient to 60.0% with a power  $(1-\beta)$  of 80.0%. The risk of  $\alpha$  in this case is  $2\alpha$  equivalent to 0.05 (5%). It is important to calculate the sample size that will be needed for this trial<sup>28</sup>. To compare the proportions in this study, the variance of the difference,  $S^2$  equals:

$p_1(1 - p_1) + p_2(1 - p_2)$  Where  $p_1$  and  $p_2$  are the proportions with response in the two groups. Thus, we have:

$$\alpha = 0.05 \Rightarrow z_{2\alpha} = 1.96$$

$$\beta = 0.20 \Rightarrow z_{\beta} = 0.84$$

$$p_1 = 0.4, \quad p_2 = 0.6\Delta = 0.2$$

Using Equation (1), we have

$$n_0 = \frac{(1.96 + 0.84)^2 \cdot 0.4(1 - 0.4) + 0.6(1 - 0.6)}{0.2^2}$$

$$= \frac{3.7632}{0.04} = 94$$

Thus, 188 patients would be needed in the study. Only 120 patients (60 in each group) were reported to have been recruited for the study due to various difficulties. By solving the general sample size formula according to  $z_{\beta}$ , we obtain<sup>16</sup>:

$$z_{\beta} = \frac{\sqrt{n}}{s} X\Delta - z_{2\alpha} \quad (2)$$

$$z_{\beta} = \frac{\sqrt{60}}{\sqrt{0.48}} X0.2 - 1.96 = 2.24 - 1.96 = 0.28$$

From Table-1 (right part),  $\beta$  is interpolated to 0.39. Therefore, the power  $1-\beta$  in this limited number of patients is now 61%. This significantly reduces the power and seriously reduces the chance of showing a significant effect. Thus, the inconclusiveness of the superiority test can be explained using the post hoc power estimate and the estimate can be used not to support a negative result of a superior study.

**Equivalence trials:** An equivalence study aims to see if a new treatment/intervention is equally effective in controlling treatment. Even if no greater therapeutic effect than a control therapy is to be expected<sup>16</sup>, equivalence trials would be useful given that the intervention is simpler with fewer side effects or is not exorbitant. Generally,  $\Delta$  should not exceed half of what

can be used in a superiority trial<sup>27</sup>. Equivalence between treatments is demonstrated if the confidence interval of the difference in effect between treatments is entirely between  $-\Delta$  and  $+\Delta$ . In this type of trial, the  $H_0$ -null hypothesis is that there is at least a difference ( $\Delta$ ), and the purpose of the test is to reject  $H_0$  in favour of  $H_A$ -alternative hypothesis that there is no difference<sup>24</sup>. The  $H_0$  and  $H_A$  therefore have a reverse role in an equivalence trial.

Although this case is a mirror image of the superiority trial. It turns out that  $\Delta$  has distinct meanings in superiority and equivalence judgments but the method of calculating the number of participants for the study is similar in the two types of trials.

**Case 2<sup>16</sup>:** In the same patients described in Case 1, an investigator wants to do an RCT to test the accuracy of pegylated interferon plus ribavirin (sustained response 40%) and another new, less expensive therapeutic strategy with fewer side effects. The investigator would calculate the sample size to determine number of patients required in the study. The power,  $1-\beta$  of the assay is considered to be 80%. The type-1-error risk ( $2\alpha$ ) is taken as 5%. The treatments are assumed equivalent if the CI estimate of the difference relative to the continuous response falls completely within  $\pm 0.10$  or  $\pm 10\%$  interval<sup>16</sup>. Therefore,  $\Delta$  is given as 0.10. Then we have:

$$\alpha = 0.05 \Rightarrow z_{2\alpha} = 1.96$$

$$\beta = 0.20 \Rightarrow z_{\beta} = 0.84$$

$$p_1 = 0.4, \quad p_2 = 0.4\Delta = 0.10$$

Using Equation (1), we have

$$n_0 = \frac{(1.96 + 0.84)^2 \cdot 0.4(1 - 0.4) + 0.4(1 - 0.4)}{0.10^2}$$

$$= 376$$

Thus, 752 patients will be the required for the study. The study was conducted and the sustained virological responses were 0.39 (39%) and 0.41 (41%) in the control and treatment groups, respectively. Difference,  $\Delta = 0.02$  with  $p > 0.50$  which is not statistically significant.

### Non inferiority trials

The purpose of a non inferiority study is to demonstrate that an intervention is no worse than the standard therapy. Therefore, in a non inferiority test the difference in efficacy (new therapy versus standard therapy) must not be smaller than  $\Delta$ <sup>16</sup>. If the lower limit for the difference in efficacy between the intervention and standard therapy is greater than  $\Delta$ , the new therapy will be proven non inferior. The location of the upper limit is not the major concern. Consequently, a non inferiority trial is conducted as a one-tailed study. Therefore, the sample size required for this study is smaller than that required for an equivalence study.

**Case 3<sup>16</sup>:** We wish to perform a similar test described in Case 2 as a non inferiority test. Therefore, the decision must be one-tailed, not the two-tailed equivalence trial. The distinction is that  $z_{\alpha}$  would be used in place of  $z_{2\alpha}$ . Given that a type-1 error rate ( $\alpha$ ) = 0.05,  $z_{\alpha}$  = 1.64 (Table-1, right side), then:

$$\begin{aligned} \alpha = 0.05 &\Rightarrow z_{\alpha} = 1.64 \\ \beta = 0.20 &\Rightarrow z_{\beta} = 0.84 \\ p_1 = 0.4, \quad p_2 = 0.6 &\Delta = 0.10 \end{aligned}$$

Using Equation (1), we have

$$\begin{aligned} n_0 &= \frac{(1.64 + 0.84)^2 \cdot 0.4(1 - 0.4) + 0.6(1 - 0.6)}{0.10^2} \\ &= 295 \end{aligned}$$

Thus, 590 patients will be required for the study. The study was conducted and the sustained virological responses were 0.39 (39.0%) and 0.42 (42.0%) in the control group and intervention, respectively. Difference,  $\Delta = 0.03$  with  $p > 0.50$  which is not statistically significant.

## Conclusion

In this review, we were able to discuss some relevant statistical concepts for understanding and estimating sample size for randomised controlled trials (RCTs). The RCT is discussed as the best way to compare the effect of treatments. This concept is demonstrated using randomised clinical trials of superiority, equivalence, and non inferiority trials with illustrative examples. The purpose of RCT is to demonstrate that an intervention is superior to a standard therapy or a placebo and can be designed and conducted as equivalence or non inferiority studies. The RCTs can also be designed and conducted as superiority studies. These studies have different objectives and differ in some methodological approaches. Knowledge of differences in methodological approaches in these studies is generally limited to the best of our knowledge and belief. This document provides an overview of the methodology of this type of testing, with particular attention to the differences in test planning, execution, analysis, and reporting. We hope that this review article will be useful to biomedical researchers and other scientists when estimating sample sizes for studies in a variety of disciplines, as the concepts and methods of calculating sample sizes discussed in this article are not limited to randomised controlled trials.

## References

- Pocock, S. J. (2013). Clinical trials: a practical approach. John Wiley & Sons.
- Armitage, P., Berry, G., & Matthews, J. N. S. (2008). Statistical methods in medical research. John Wiley & Sons.
- Kirby, A., Gebiski, V., & Keech, A. C. (2002). Determining the sample size in a clinical trial. *Medical Journal of Australia*, 177(5), 256-257.
- Campbell, M. J., Machin, D., & Walters, S. J. (2010). Medical statistics: A textbook for the health sciences. John Wiley & Sons.
- Chan, Y. H. (2003). Randomised controlled trials (RCTs)-sample size: the magic number?. *Singapore medical journal*, 44(4), 172-174.
- Sharma, S. K., Mudgal, S. K., Thakur, K., & Gaur, R. (2020). How to calculate sample size for observational and experimental nursing research studies. *National Journal of Physiology, Pharmacy and Pharmacology*, 10(1), 1-8.
- Wittes, J. (2002). Sample Size Calculations for Randomized Controlled Trials. *Epidemiol Rev.*, 24(1), 39-53.
- Petrie, A. & Sabin, C. (2019). Medical statistics at a glance. John Wiley & Sons.
- Lombardi, R. (2014). Designing randomized clinical trials in surgery. *British Journal of Surgery*, 101(4), 293-295.
- Fayers, P.M., Cuschieri, A., Fielding, J., Craven, J., Uscinska, B. and Freedman, L.S. (2000). Sample Size Calculation for Clinical Trials: The Impact of Clinician Beliefs. *British Journal of Cancer*, 82(1), 213-219.
- Jacob Cohen (1977). Statistical Power Analysis for the Behavioral Sciences. 2<sup>nd</sup> ed. London: Academic Press.
- Malterud, K., Siersma, V. D., & Guassora, A. D. (2016). Sample size in qualitative interview studies: guided by information power. *Qualitative health research*, 26(13), 1753-1760.
- Fayers, P. M., & Machin, D. (1995). Sample size: how many patients are necessary?. *British Journal of Cancer*, 72(1), 1-9.
- Kaji, A. H., & Lewis, R. J. (2010). Are we looking for superiority, equivalence, or noninferiority? Asking the right question and answering it correctly. *Annals of Emergency Medicine*, 55(5), 408-411.
- Ward, M. M. (2007). Primer: measuring the effects of treatment in clinical trials. *Nature Clinical Practice Rheumatology*, 3(5), 291-297.
- Erik, C. (2007). Methodology of Superiority vs. Equivalence Trials and Non-inferiority Trials. *Journal of Hepatology, Elsevier*, 46, 947-954
- Makuch, R., & Simon, R. (1978). Sample Size Requirements for Evaluating a Conservative Therapy. *Cancer Treatment Reports*, 62(7), 1037-1040.
- Sander, A., Rauch, G. & Kieser, M. (2017). Blinded sample size recalculation in clinical trials with binary composite

- endpoints. *Journal of Biopharmaceutical Statistics*, 27(4), 705-715.
19. Kieser, M., & Hauschke, D. (1999). Approximate sample sizes for testing hypotheses about the ratio and difference of two means. *Journal of Biopharmaceutical Statistics*, 9(4), 641-650.
20. Fleiss, J. L. (1992). General design issues in efficacy, equivalency and superiority trials. *Journal of periodontal research*, 27(4), 306-313.
21. Garrett, A. D. (2003). Therapeutic equivalence: fallacies and falsification. *Statistics in Medicine*, 22(5), 741-762.
22. Blackwelder, W. C. (1982). Proving the null hypothesis in clinical trials. *Controlled clinical trials*, 3(4), 345-353.
23. Greene, W. L., Concato, J., & Feinstein, A. R. (2000). Claims of equivalence in medical research: are they supported by the evidence?. *Annals of Internal Medicine*, 132(9), 715-722.
24. Costa, L. J., Xavier, A. C. G., & Del Giglio, A. (2004). Negative results in cancer clinical trials-equivalence or poor accrual?. *Controlled Clinical Trials*, 25(5), 525-533.
25. Dimick, J. B., Diener-West, M., & Lipsett, P. A. (2001). Negative results of randomized clinical trials published in the surgical literature: equivalency or error?. *Archives of surgery*, 136(7), 796-800.
26. Detsky, A. S., & Sackett, D. L. (1985). When was a 'negative' clinical trial big enough?: how many patients you needed depends on what you found. *Archives of Internal Medicine*, 145(4), 709-712.
27. Jones, B., Jarvis, P., Lewis, J. A., & Ebbutt, A. F. (1996). Trials to assess equivalence: the importance of rigorous methods. *BMJ*, 313(7048), 36-39.
28. Ryan, T. P. (2013). *Sample size determination and power*. John Wiley & Sons.