Chuka University
info@chuka.ac.ke | 020 231 0512 / 020 231 0518

GENERAL OVERVIEW OF SAMPLE SIZE ESTIMATION FOR RANDOMIZED CONTROLLED CLINICAL TRIALS

Obare, D.M.*, Gladys G. Njoroge, and Muraya, M.M.
Department of Physical Sciences, Chuka University, P. O. Box 109-60400, Chuka, Kenya Corresponding author email: obaredominic87@gmail.com
moses.muraya@chuka.ac.ke

How to cite:
Obare, D. M., Njoroge, G. G. and Muraya, M. M. (2021). General overview of sample size estimation for randomized controlled clinical trials. *In: Isutsa, D. K. (Ed.). Proceedings of the 7th International Research Conference held in Chuka University from 3rd to 4th December 2020, Chuka, Kenya, p. 537-546*

ABSTRACT
Calculation of the minimum sample size needed to meet the primary study objective is a key feature of the design of any clinical trial. The other reason a priori sample size determination is to limit participant harm or loss of clinical benefit to as few study participants as possible. This article generally reviews the basic principles that determine an appropriate sample size and provides methods for its calculation in some simple, yet common, cases. Sample size is closely tied to statistical power, which is the ability of a study to enable detection of a statistically significant difference when there truly is one. A trade-off exists between a feasible sample size and adequate statistical power. **Keywords:** Clinical trial, Sample size, Statistical power, Randomization, Review, Participant.

INTRODUCTION
The clinical trial is the most definitive tool for evaluation of the applicability of a clinical research. It represents "a key research activity with the potential to improve the quality of health care and control costs through careful comparison of alternative treatments (Freiman, Chalmers, Smith and Kuebler, 1978). A properly planned and executed clinical trial is the best experimental technique for assessing the effectiveness of an intervention. It also contributes to the identification of possible harms (Friedman et al., 2015). Clinical trials are voluntary prospective studies conducted in human beings and designed to answer specific questions about the safety or effectiveness of drugs, vaccines, other therapies or new interventions (Society for clinical trials, 2006).

A proper and sufficiently large randomized clinical trials are the best way to ascertain which interventions are effective and safe in order to improve public health (Pezzullo, 2014). To determine the correct number of participants to involve is key to a meaningful clinical trial. A large number of participants will make a clinical trial unfeasible while a small number will make the trial have a very low statistical power and its findings will be considered of less impact or the intervention is not effective. To avoid this challenge a researcher needs to determine the correct number of participants to be recruited in a clinical trial. The number of participants in a clinical trial is known as the sample or the study population (Friedman et al., 2015). The study population is determined in the planning phase when developing the study protocol. This is important since, in claiming an intervention is or is not effective it is essential to describe the type of participants on which the intervention will be tested. These includes, specification of criteria for eligibility and description of who actually should be enrolled.

In reporting and assessing the findings of a clinical trial the researcher needs to say what population was studied and how they were selected. This is because, (1) if an intervention is shown to be successful or unsuccessful, the medical and scientific communities must know to what population the findings refer (Campbell et al., 1995), (ii) knowledge of the study population helps other investigators asses the study's merit and appropriateness (Warner, 1995), (iii) in order for other investigators to be able to replicate the study, they need data descriptive of those enrolled. However, this paper does not discuss into detail all aspects of the study population in clinical trials. Well elaborate description and other details of study population can be obtained from (Duley & Farrell, 2002, Cleophas *et al*., 2013, Friedman *et al*., 2015, Piantadosi, 2017). The scope of this paper is limited to general statistical methods used to estimate sample size in randomized clinical trials. It brings together different statistical methods that have been utilized to estimate sample sizes in clinical trials. It

highlights the shortcomings of these methods in estimating the sample size for clinical trials. This paper focuses only on estimation of sample size of clinical trials with single primary response variable, multiple primary response variables.

DISCUSSION

The size of the study should be considered early in the planning phase when laying down the study protocol. In most cases, no formal sample size is determined. The number of participants available to the investigators during a particular period of time will dictate the size of the study. If the study requires more participants, enrollment continues as the follow-up continues. If clinical trials do not consider the sample size requirements, they turn out to lack the statistical power or ability to prove that intervention effects are of clinical importance.

In a review of 71 published randomized "negative" controlled clinical trials by Freaman and colleagues in 1978. It found out that 67 of the trials had greater than 10% risk of missing a true 25% therapeutic improvement, and with the same risk, 50 of the trials could have missed a 50% improvement. Estimates of 90% confidence intervals for the true improvements in each trial showed that in 57 of these "negative" trials a potential of 25% improvement was possible, and 34 of the trials showed a potential of 50% improvement. This was attributed to small sample sizes and needs more attention in the planning of clinical trials (Freiman, Chalmers, Smith and Kuebler, 1978).

This situation remained the same even in 1994, 383 randomized controlled trials(RCTS) with negative results were reviewed on the basis of the power to detect 25% and 50% relative difference. 27% were classified as having negative results. Only 16% and 36% hard sufficient statistical power. In general, only 32% of the trials with negative results reported sample size estimation. Most trials with negative results did not have large enough sizes to detect a 25% or 50% relative difference (Moher et al., 1994). The danger that underlies this trend of conducting unethical underpowered clinical trials is that interventions that could be beneficial are discarded without adequate testing and may never be considered again (Friedman et al., 2015). Hitherto, many studies do contain appropriate sample size estimates, however, after many years of critical review majority are still too small (Chan AW, & Altman DG. 2005, Halpern SD, Karlawish JH, & Berlin JA, 2002, Friedman et al., 2015).

Basic Statistical Theory

In order to understand how to determine a meaningful sample size for clinical trials, there are basic fundamental principles we need to familiarize with. These include hypothesis testing, significance level and statistical power (Gueyffier & Boissel, 1998, Jia & Lynn, 2015, Gong *et al*., 2000, Fisher & Belle, 1997). This will be helpful to those with no background of the basic statistical concepts.

Clinical trials can be of one intervention group and one control group (Friedman et al.,2015, Day SJ& Graham DF, 1991). In estimating the sample size for clinical trials, the primary response variable used to judge the effectiveness of intervention should be identified. A primary response corresponds to a primary question that the investigators and sponsors are most interested in answering. The investigators can focus on a single primary response variable or multiple responses, this largely depends on what the investigators are interested in and the complexity of the clinical trial. The sample size largely is based on the primary questions and their corresponding responses (Cutler *et al*., 1966, Friedman *et al*., 2015). Generally, there are 3 different responses or outcomes; (i) dichotomous/binary response variable (ii) continuous response variable and (iii) time to failure/occurrence of a clinical event (Koch *et al*., 2015).

Single Primary Response Variable.

The single primary response variable is an outcome obtained at the end of the follow-up of a clinical trial (Clarke, 2007, Donner, 1984). These single primary responses can either be dichotomous, continuous or time to failure (survival) responses. For the dichotomous response variables, let the event rates in the intervention group be (P_I) and the control group be (P_C) . For the continuous responses let the true but unknown mean level in the intervention group be (μ_I) and the mean level in the control group be (μ_C). For survival rate outcomes the hazard rate ( ) is often used as an event rate estimator (Donner 1984, Friedman *et al*., 2015).

Researchers and investigators do not know the true values of the event rates. Clinical trials give only estimates of the event rates, $\hat{P}_I$ and $\hat{P}_C$. The investigator in a clinical trial is only interested with whether or not a true difference exists between the event rates of participants in the two or multiple groups. Conventionally, this is done through the

null hypothesis, denoted as H_O; which states that no difference exists between the true event rates. The

investigator needs to test H_O and decide whether to reject it or not. The null hypothesis is assumed to be true until proven otherwise. (Schneider, 1981).

If the event rates in a clinical trial analysis are large enough by chance alone, then it means the investigator might reject the null hypothesis incorrectly. This is known as false positive finding or Type I error. This error should be controlled or minimized as far as possible. The probability of Type I error is known as the significance level, denoted as α. Given H_O is true, there is the probability of observing differences as large as or large than the

difference known as "P-value" normally denoted as . The statistical decision is reject H_O, if ≤α. α is chosen

arbitrarily, the ones used and accepted in the literature are 1%, 2.5% and 5%. (Schneider, 1981, Blackwelder, 1981, Giangregorio & Cook, 2009, Ng, 1995).

When the null hypothesis is not true, then another hypothesis called the alternative hypothesis, denoted by H_A must be true. The true difference between the event rates is a value $\delta \neq 0$. Sometimes the observed difference between event rates can be very small by chance alone even if the alternative hypothesis is true. Due to this small observed differences, the researcher could fail to reject the H_O even when it is not true. This will result to a Type II error, or a false negative. The probability of a type II error is denoted by . 1 – is the probability of correctly rejecting H_O . 1 – is known as the statistical power of the study. The power gives the potential of the study to find the true differences of various values of δ (Gueyffier & Boissel, 1998, Ghosh, 2002 "Will my clinical study be a success? The Concept of Statistical Power", 2020).

In controlled randomized clinical trials, it is assumed that randomization will allocate an equal number ( ) of participants to each group, that is intervention and control groups. This ensures a more powerful design than unequal allocation (Trachtman & Caplan, 2018). Equal allocation is usually easier to implement therefore it is more frequently used strategy because of its simplicity in the analysis (Hey & Kimmelman, 2020).

From classical statistical theory, the researcher must make a decision if he is interested with differences in one direction only (one-sided test) or in differences in either direction (two-sided test). If one-sided test of hypothesis is chosen, mostly from the literature the significance level should be half of what the investigator would use for a two- sided test. For example, if 5% is the two-sided significance then, 2.5% would be used for the one-sided test.

In a clinical trial the total sample size is 2 (per arm). Clearly, the sample size estimation is a function of δ, α and , 1 – . Any change to either these parameters will result in a change in sample size 2 (per arm). If δ decreases, then the sample size should be large enough to guarantee a high probability of finding the difference. If the calculated sample is unrealistically large than it can be obtained, then one or more of the parameters in the design may need to be re-adjusted.

Since the significance level is usually fixed at 5%,2.5% or 1%, the researcher should reconsider the value selected for δ and increase it, or keep δ the same and settle for a less powerful study. If neither of these alternatives is satisfactory, serious consideration should be given to abandoning the trial (Cleophas et al., 2013, Clinical trials, 2020).

Sample Size estimation for Dichotomous/Binary Response Variables *Two Independent Samples*

The primary response variable is the occurrence of an event over some fixed period of time. The sample size calculation should be based on the specific test statistic that will be employed to compare the outcomes. The null hypothesis H_0 (P_C - P_I =0) (no difference) is compared to an alternative hypothesis H_A (P_C - $P_I \neq 0$) (P, 2020). The estimates of P_C and P_I are $\hat{P}_I, \hat{P}_C$ respectively, where $\hat{P}_I = \frac{r_I}{N_I}$ and $\hat{P}_C = \frac{r_C}{N_C}$ with r_I and r_C being the number of events in the intervention arm and control arm respectively while N_I and N_C being the number of

participants in each group. The sample size required for the design to have a significance level α and a power of 1 − to detect true differences of at least δ between the event rates P_I and P_C can be expressed by the formula (Donner & Makuch, 1985, Friedman et al., 2015).

$$N = [\{Z_\alpha \sqrt{\bar{p}(1-\bar{p})} + Z_\beta \sqrt{p_C(1-p_C) + P_I(1-P_I)}\} / (p_2 - p_1)^2] \quad [1]$$

Where = sample size per arm (participants/group) with p $= \frac{1}{2}\{_C p \quad p_I +$ }, Z_α is the critical value

which corresponds to the significance level and Z_β is the value of the standard normal value not exceeded with probability β. Z_β corresponds to the power 1 − (e.g. if 1 − =0.90 Z_β =1.282). Values of Z_α and Z_β are given in generated tables which can be downloaded from online resources (Rochon, 2005).

When by $N_I = N_C$ i.e. when the two arm groups are of equal size. An alternative to the above formula is given

$$_2 = \frac{1}{2} \quad _{[(+)2} \quad p(1-p)/(p_C - p_I)_2 \quad] \qquad [2]$$

These two formulas give approximately the same answer and either may be used for the typical clinical.

By a matter of fact, participants in clinical trials do not always fully adhere with the intervention being tested. Some fraction (R_O) of participants on intervention drop-out of the intervention due to some reasons and some other fraction (R_I) drop-in and start following the intervention. The assumption here is that, participants who drop-out respond as if they had been on control and those who drop-in respond as if they had been on intervention, then the sample size adjustment is the same as for the case of proportions. That is, the adjusted sample size N* is a function of the drop-out rate, the drop-in rate, and the sample size N for a study with fully compliant participants: (Friedman *et al*., 2015, Cleophas et al., 2013)

$N^* = N/(1-R_O-R_I)^2$ [3]

Most of the methods in statistical estimation of sample size in clinical trials have not incorporated this aspect of drop-in and drop-out of participants. In this paper this aspect will be considered and it will be included in the final formula for estimation of the sample size. The two formulas above precisely should be as follows, [1] becomes:

$$N = [\{Z_\alpha \quad p\sqrt{(1-p)} + Z_\beta \quad \sqrt{\quad} C(1-p_C) + P_I(1-P_I)\}/(p_2 \quad p_I)^2 \quad]/(1-R_O-R_I)^2$$

And [2] becomes

$$_2 = \frac{1}{2} \quad _{[(+)2} \quad p(1-p)/(p_C \quad p_I)^2 \quad]/(1-R_O-R_I)^2$$

Where $(1-R_O-R_I)^2$ is the correction factor for drop in and drop-out of participants.

Sample Size Estimation for Continuous Response

Variables *Two Independent Samples*

Continuous response variables include blood pressure, Spiro metric measures, neuropsychological scores, level of a serum component and length of hospitalization just to mention a few. These variables are measured during a clinical trial in order to provide data for statistical analysis.

Let the primary response variable be denoted as x, is continuous with N_I and N_C participantsrandomized to the

intervention group and control group respectively. Let us assume that the variable is normally distributed with mean μ and variance σ^2. True levels of and for the intervention and control groups are unknown, but it is assumed that σ^2 is known. Normally, σ^2 is unknown and must be estimated from some sample data. If the data set used is large enough, the estimate of σ^2 can be used in place of the true σ^2 (Chow, 2011). But this is not easy to evaluate hence it leaves investigators at crossroads on the efficacy of the estimates to be adopted.

The null hypothesis is H_0 : δ= μ_c- μ_i = 0 (no significant difference) and the two-sided alternative hypothesis is H_A :

$\delta = \mu_c - \mu_i \neq 0$ (there is a difference) (Bristol, 1992). If the variance is known, the test statistic is:

$$Z = (\bar{X}_C - \bar{X}_I)/\sigma \left(1/N_C + 1/N_I\right)^{\frac{1}{2}} \quad [4]$$

Where

$\bar{X}_I$ and $\bar{X}_C$ represent mean levels observed in intervention and control arms (Friedman *et al*., 2015)

The sample size can be estimated as

$$N=\{(Z_\alpha+Z_\beta)^2 \sigma^2 / \delta^2\} /(1-R_O -R_I)^2 \quad [5]$$

Where the $(1-R_O -R_I)^2$ is the correction factor for the drop-in and drop-out of participants.

Paired Data

In some clinical trials, paired outcome data may increase power for detecting differences because individual or within participant variation is reduced. Trial participants are normally assessed at baseline and at the end of follow- up. Let's assume that Δ_C and Δ_I represent the true, but unknown levels of change from baseline to some later point in the trial for the control and intervention groups, respectively (Gauderman, 1992).

Estimates of Δ_C and Δ_I would be

$$\bar{d}_C = \bar{X}_{C1} - \bar{X}_{C2} \text{ and } \bar{d}_I = \bar{X}_{I1} - \bar{X}_{I2}.$$

These represent the differences in mean levels of the response variable at two points for each group. The investigator tests H_0: $\Delta_C - \Delta_I = 0$ against H_A: $\Delta_C - \Delta_I \neq 0$. Using δ and σ^2_Δ, as defined in this manner, the previous sample size formula for two independent samples are applicable (Rosner, 1982). The total sample size per arm N can be estimated as (Yelland et al., 2017)

$$N= \{ (Z_\alpha +Z_\beta)^2 (1-\rho) \sigma^2_\Delta /\delta^2 \} /(1-R_O -R_I)^2 \quad [6]$$

Where the $(1-R_O -R_I)^2$ is the correction factor for the drop-in and drop-out of participants (Friedman *et al*., 2015).

Sample Size Estimation for Repeated Measures clinical trials

These are clinical trials whose primary responses are a continuous response variable measured at each follow-up visit unlike the methods presented earlier that consider the sample size calculation for trials where only a baseline and a final visit are used to estimate the effect of intervention (Lu et al., 2008). This approach is useful in thinking about how many participants, how many responses per individual, and the time that they should be taken, are needed. We assume that the change in response variable is a linear function of time, then the rate of change is summarized by a slope. This model is fit to each participant's data by the standard least squares method and the estimated slope is used to summarize the participant's experience (Herring, 2013). The investigator must be concerned about the frequency of the measurement and the duration of the observation period. The observed measurement x can be expressed as = + + , where = , = , = , and error represents the deviation of the observed measurement from a regression line (Berlin & Ness, 1996, Chow & Liu, 2013). The error is due to measurement variability, biological variability or the nonlinearity of the true underlying relationship. The error is equally distributed around 0 and have a variability denoted as σ^2_{erro}. Let us assume that σ^2_{erro} is the same for every participant. The investigator will evaluate intervention effectiveness by comparing the average slope in one group with the average slope in another group. The slope variability reflects the effectiveness of the intervention or control. The amount of variability of slopes over participants is denoted as σ^2_b.

If D represents the total time duration for each participant and P represents the number of equally spaced measurements, σ^2_b can be expressed as:

$$\sigma^2_b = \sigma^2_B + \{12(P-1)\sigma^2_{erro} /(D^2P(P+1))\} \quad [7]$$

where σ^2_B is the component of variance attributable to differences in participants' slope as opposed to measurement error and lack of a linear fit (Friedman *et al*., 2015, Fairclough, 2010). The sample size required to detect difference δ between the average rates of change in the two groups is given by:

$$N = [4(Z_\alpha +Z_\beta)^2/ \delta^2][\sigma^2_B + \{12(P-1)\sigma^2_{erro} / (D^2P(P+1))\}]* \frac{1}{(1-R_O -R_I)^2} \quad [8]$$

B error —
2

Where the $(1-R_O-R_I)^2$ is the correction factor for the drop-in and drop-out of participants.

Sample Size Estimation for "Time to Failure"

In time to failure outcomes we employ life tables or survival analysis methods. Survival curves for the groups are compared to measure the effectiveness of an intervention. Non-parametric models are commonly used and this avoids any assumption on mathematical model of the survival curve. However, for sample size estimation, some assumptions are vital. We assume that the survival curve, S(t), follows an exponential distribution, ( ) = − = (−) where λ is the hazard rate. Using this model, survival curves are totally characterized by λ. Thus, the survival curves from a control and an intervention group can be compared by testing 0: = . An estimate of λ is obtained as the inverse of the mean survival time. If the median survival time, T_M, is known, the hazard rate λ may also be estimated by -ln (0.5)/T_M. Sample size formulations have been considered by several investigators. One simple formula is given by Lachin, 1981.

where N is the size of the sample in each group (per arm) and Z_α and Z_β are defined as above.

The method just described above assumes that all participants will be followed to the event. Clinical trials with a survival outcome are terminated at time T before all participants have had an event. For those still event-free, the time to event is said to be censored at time T. (Lachin, 1981).

For this situation, alternative formula is: = (+)2 [( ) + ( )]/ (−)2 Where ( ) = 2/(1 − −) and where ( ) ( ) are defined by replacing λ with , respectively.

Further models can be obtained in George & Desu, 1974 and Lachin, 1981. The methods presented have some limitations in estimating the sample size for clinical trials. They are presented with the view that they can be used to estimate the sample size for most of the clinical trials. The methods take into consideration so many assumptions which raise concerns on the power of the study generated from those estimates. The event rates are based on estimations and approximations from previous studies of the same groups or people.

This is a challenge since obtaining such information is difficult in many cases. Also the event rates evidenced in the literature are based on small sample sizes with very low statistical power which will make the trials unethical if the same approximations are utilized in future studies. The statistical methods above do not specify at what phase they can be utilized, as we are aware depending with the intervention in question the clinical trial is a complex task that has a well-defined protocol which must be adhered to. Sample sizes of clinical trials are specific and usually given a range at each and every phase. It is imperative for any method presented for sample size estimation to address this issues. The methods have addressed the issue of drop-in and drop-out of participants but what if a participant did not drop-out completely i.e. if the participants miss interventions or doses within the follow-up period.

Sample size Estimation for Bioequivalence Trials

When an effective intervention for example a drug product has been established and is considered the standard, efforts are made by researchers to develop new interventions that are less expensive, have less side effects, or have less adverse impact on an individual's general quality of life (Proschan, 2009). This kind of studies are common in the field of pharmaceutical industry where a drug developed by one company may be tested against an established and marketed drug by another company. These kind of studies or trials have positive controls or commonly known as non-inferiority designs (Classen, 2004).

Currently, there is no statistical model for designing non-inferiority trials to demonstrate complete equivalence. It is not statistically possible to show that there is no difference between two interventions in this kind of trials. If the investigator finds evidence not to reject the null hypothesis is not sufficient to claim that the two interventions are equal but a lack of adequate evidence to show they are different. (Qu & Zheng, 2003). In bioequivalence trials we need to specify δ i.e., interventions with differences which are less than might be considered equally effective or non-inferior. More discussion on non-inferiority trial can be obtained from (Herchuelz, 1996, Patterson & Jones, n.d., Hauschke et al., 2007). Specification for δ is a difficult task, it is left to the experts to determine depending with the level of tolerance. For dichotomous responses we can assume interventions to be equal to P (i.e. $P=P_C=P_I$.). The sample size can be estimated as:

Where ,are defined as above (Makuchi and Simon, 1978)

As discussed in Friedman *et al*., 2015, specifying δ, is a fundamental part of the design and sample size estimations of all equivalence and non-inferiority trials. Trials should be sufficiently large, with enough power to address properly the

questions about equivalence studies. Bioequivalence trials are complex in nature due to the importance they carry in medical interventions. Apart from the method presented above for sample estimation other methods have been proposed (Chow, 2011, Pharmacokinetic and bioequivalence studies, 1991, Jones & Kenward, 2003).

Usually, bioequivalence studies are conducted under cross-over designs or parallel designs with raw data or log-transformed data. Researchers have to consider intra-subject and inter-subject variation when estimating sample size in cross-over bioequivalence trials. Parallel designs are not very complicated to implement but considerations should be taken not to carry out a wasteful clinical trial. Another limitation is the decision by the use of either log- transformed or raw data, which one should be adopted without much assumptions and still give desired results. More detailed discussion on sample size estimation for bioequivalence trials can be obtained from Hauschke, 2002.

Sample size estimation for clinical trials with multiple end-points

Most clinical trials have more than one primary question and a single primary response variable. Having a single primary question and a single primary response variable is advantageous because of its simplicity in terms of design and even execution, but this comes in handy with numerous assumptions which sometimes reduces the efficacy of the study. Where investigators cannot agree on which outcome is most important on judging the intervention, it results to multiple outcomes. This is true by the fact that effects of interventions are multi-dimensional. Even though multiple primary endpoint design is challenging, it is also advantageous since it captures more complete characterization of the intervention effects and provide more informative intervention comparisons.

For these reasons, use of more than a single primary endpoint has become a common design in clinical trials for disease areas like oncology, infectious diseases, cardiovascular disease and bioequivalence trials in pharmaceutical industries. For example, a clinical trial involving participants with pulmonary embolism (Urokinase Pulmonary Embolism Trial Study Group, 1974) employed 3 methods of determining a drug's ability to resolve emboli. They were; lung scanning, arteriography, and hemodynamic studies.

Estimation of sample size for such clinical trials is non-trivial. The resulting need for new approaches to the design and analysis of clinical trials has been proposed (Dmitrienko et al., 2010; Gong et al., 2000, Hung and Wang 2009; Offen *et al*., 2007). Controlling type I and type II error in clinical trials involving multiple endpoints is also non- trivial. Gordoba et al., 2010 mentioned that correlation among the multiple endpoints should be considered when estimating sample size. Correlation in multiple endpoint trials is usually unknown and therefore must be estimated with external data. They proposed that multiple endpoints can be treated as a single composite endpoint. This will effectively reduce the problem to a single dimension hence simplifying the design to avoid the multiplicity issues regarding multiple endpoints. Creation and interpretation of a composite endpoint is challenging when the treatment effects vary across components with very different clinical importance.

Hamasaki et al. (2012) discussed the sample size estimation for trials with multiple risk ratios and odds ratios as primary contrasts. Sozu et al. (2010, 2011) presented the overall power and sample size calculations in bioequivalence clinical trials with co-primary binary endpoints assuming that the binary endpoints are jointly distributed as a multivariate Bernoulli distribution. They found out that there are significant challenges in estimation of the correlation due to the multiplicity of endpoints resulting to restrictions on the correlation. Song (2009) discussed sample size calculations with co-primary binary endpoints in non-inferiority clinical trials, but never mentioned anything to do with restrictions on the correlation.

Xiong et al. (2005) discussed power and sample size for clinical trials with two co-primary continuous endpoints with the assumption that the two endpoints are bivariate normally distributed and their variance-covariance matrix is known. Sozu et al. (2006) further discussed this and extended it to continuous endpoints with the assumption that the variance-covariance matrix is unknown using the Wishart distribution. Sozu et al. (2011) discussed extensions to more than two continuous endpoints for both known and unknown variances. Julious and McIntyre (2012) presented three methods of sample size estimation of clinical trials involving multiple comparisons. Since the testing procedure for co-primary endpoints may be conservative, the methods may result into large and impractical sample sizes. Patel (1991), Stein et al. (2007) and Kordzakhia et al. (2010) discussed methods to control the type I error rate in order to address some of the issues above. However, these methods can lead to smaller sample sizes which is unethical, and other technical issues.

Sozu, Sugimoto, Hamasaki and Evans, 2015 discussed sample size determination in clinical trials with multiple endpoints. They suggested that correlation among endpoints should be incorporated into power and sample size estimation. However, this leads to reduction of sample size. They presented methods to calculate sample sizes for continuous co-primary endpoints with known and unknown variances, binary co-primary endpoints and continuous primary endpoints. They also

noted that these methods were only designed to address specific scenarios of multiple endpoints trials. Before their utilization correct assumptions must be taken to avoid difficulties. The methods discussed did not take into consideration clinical trials with more than two interventions, clinical trials with mixed endpoints and group sequential clinical trials. Perhaps with correct assumptions the methods can be extended to estimate sample sizes for other clinical trial designs. These methods require considerable mathematical sophistication and knowledge of programming techniques and this effectively limits their application in practice.

CONCLUSIONS

Sample size estimation in clinical trials must be addressed before the start of a randomized, controlled trial. It is crucial to know in advance the likelihood of finding valid conclusions from the population assessed. It may be acceptable to subject a patient to a chance of a less than ideal treatment, or to the psychological stress of being a 'subject', if there is a chance of a valid scientific outcome, but it is unethical to conduct a study the design of which cannot make valid conclusions. Knowledge of the requirement of sample size is necessary in the planning of a prospective study, and such information should encourage investigators into engaging in multicenter trials. The performance of power calculations specifically demands that the minimal effect of interest is established and calls attention to important details which may otherwise be overlooked.

A proper design and appropriate statistical analysis are essential to validity of all quantitative clinical research. A Type-I error is better known than a type II error and reviewers and readers are more cognizant of -values when authors conclude that significant differences between groups are found. Equal scrutiny is required when authors decide that there is no statistically significant difference. In this age of limited resources and tight budgets, physicians may be forced to employ cheapest methods, especially if choices are thought to be similar. It is important to consider associations among endpoints into sample size calculation when the endpoints are correlated and the effect sizes are approximately equal among the endpoints. In bioequivalence trials, it is important that investigators do not erroneously label two treatments as equivalent, when it has merely been shown that the differences were not statistically significant. All clinical studies should be based on appropriate calculations of sample size.

REFERENCES

Berlin, J., & Ness, R. (1996). Randomized clinical trials in the presence of diagnostic uncertainty: Implications for measures of efficacy and sample size. *Controlled Clinical Trials*, *17*(3), 191-200.

Blackwelder, W. (1981). "Proving the Null Hypothesis" in clinical trials. *Controlled Clinical Trials*, *2*(1), 67. Bristol, D. (1992). Sample size methodology. *Controlled Clinical Trials*, *13*(6), 519-520. https://doi.org/10.1016/0197-2456(92)90211-h

Campbell, M., Julious, S., & Altman, D. (1995). *Estimating sample sizes for binary, ordered categorical, and continuous outcomes in two group comparisons*. *BMJ*, *311*(7013), 1145-1148.

Chan AW, Altman DG. Epidemiology and reporting of randomized trials published in PubMed journals. Lancet 2005; 365:1159–1162

Chow, S. (2011). *Controversial statistical issues in clinical trials*. CRC Press.

Chow, S. (2011). Sample size calculations for clinical trials. *Wiley Interdisciplinary Reviews: Computational Statistics*, *3*(5), 414-427. https://doi.org/10.1002/wics.155

Chow, S., & Liu, J. (2013). *Design and analysis of clinical trials*. Wiley.

Clarke, M. (2007). Standardizing outcomes for clinical trials and systematic reviews. *Trials*, *8*(1). https://doi.org/10.1186/1745-6215-8-39

Classen, H. (2004). *European quality standards applied to bioequivalence trials on Turkish generic products*. ECV.

Cleophas, T., Zwinderman, A., Cleophas, T., Victor, N., & Jensen, K. (2013). *Statistics Applied to Clinical Trials*. Springer Netherlands.

Cordoba G, Schwartz L, Woloshin S, Bae H, Gotzsche PC (2010) Definition, reporting, and interpretation of composite outcomes in clinical trials: systematic review. Br Med J 341:c3920

Cutler SJ, Greenhouse SW, Cornfield J, Schneiderman MA. The role of hypothesis testing in clinical trials: biometrics seminar. J Chronic Dis 1966; 19:857–882.

Day SJ, Graham DF. Sample size estimation for comparing two or more treatment groups in clinical trials. Stat Med 1991; 10:33–43

Dimasi, Joseph A; Grabowski, Henry G; Hansen, Ronald W (2016). "Innovation in the pharmaceutical industry: New estimates of R&D costs". Journal of Health Economics. ***47****:20–33.*

Dmitrienko A, Tamhane AC, BretzF (2010) Multiple testing problems in pharmaceutical statistics. Chapman & Hall/CRC, Boca Raton

Donner, B., & Makuch, R. (1985). Approaches to sample size estimation in the design of clinical trials—a review. *Statistics in Medicine*, *4*(2), 247-247. https://doi.org/10.1002/sim.4780040215 Duley, L., & Farrell, B. (2002). *Clinical trials*. BMJ Books.

Fairclough, D. (2010). *Design and Analysis of Quality of Life Studies in Clinical Trials, Second Edition*. Chapman & Hall/CRC.
Fisher, L., & Belle, G. (1997). Biostatistics: A Methodology for the Health Sciences. *Biometrics*, *53*(3), 1182.
Freiman JA, Chalmers TC, Smith H Jr, Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. Survey of 71 "negative" trials. The New England Journal of Medicine. 1978 Sep;299(13):690-694. DOI: 10.1056/nejm197809282991304. Friedman, L. (2015). *Fundamentals of Clinical Trials*. Springer.
Gauderman, W. (1992). Sample Size Calculations for Ophthalmologic Studies. *Archives of Ophthalmology*, *110*:690
George, S., & Desu, M. (1974). Planning the size and duration of a clinical trial studying the time to some critical event. *Journal of Chronic Diseases*, *27*(1-2), 15-24. https://doi.org/10.1016/0021-9681(74)90004-6
Ghosh, D. 2002. Statistical aspects of the design and analysis of clinical trials. *Controlled Clinical Trials* 23:299– 300
Giangregorio, L., & Cook, R. (2009). Hypothesis testing in clinical and basic science research. *Transfusion*, *50*(9), 1878-1880. https://doi.org/10.1111/j.1537-2995.2009.02536.x
Gong J, Pinheiro JC, DeMets DL (2000) Estimating significance level and power comparisons for testing multiple endpoints in clinical trials. Control Clin Trials 21:323–329
Gong, J., Pinheiro, J., & DeMets, D. (2000). Estimating Significance Level and Power Comparisons for Testing Multiple Endpoints in Clinical Trials. *Controlled Clinical Trials*, *21*(4), 313-329.
Gueyffier, F., & Boissel, J. (1998). Power, clinical and statistical significance of the search of interactions between treatment effect and patient profiles: A practical approach. *Controlled Clinical Trials*, *19*(3), S90.
Gueyffier, F., & Boissel, J. (1998). Power, clinical and statistical significance of the search of interactions between treatment effect and patient profiles: A practical approach. *Controlled Clinical Trials*, *19*(3), S90.
Hamasaki T, Sugimoto T, Evans SR, Sozu T (2013) Sample size determination for clinical trials with co-primary outcomes: exponential event-times. Pharma Stat 12:28–34
Hauschke, D., 2002. Journal of Pharmacokinetics and Pharmacodynamics, 29(1), pp.89-94.
Hauschke, D., Steinijans, V., & Pigeot, I. (2007). *Bioequivalence studies in drug development*. John Wiley & Sonx.
Herchuelz, A. (1996). Bioequivalence assessment and the conduct of bioequivalence trials: A European point of view. *European Journal of Drug Metabolism and Pharmacokinetics*, *21*(2), 149-152.
Herring, A. (2013). Applied Longitudinal Analysis, 2nd Edition, by Garrett M. Fitzmaurice, Nan M. Laird, and James H. Ware, John Wiley & Sons, 2011. *Journal of Biopharmaceutical Statistics*, *23*(4), 940-941. https://doi.org/10.1080/10543406.2013.789817
Hey, S., & Kimmelman, J. (2020). *The questionable use of unequal allocation in confirmatory trials*. Retrieved 16 May 2020, from.
Hung H M, Wang SJ (2009) Some controversial multiple testing problems in regulatory applications. J Biopharm Stat 19:1–11
Jia, B., & Lynn, H. (2015). A sample size planning approach that considers both statistical significance and clinical significance. *Trials*, *16*(1). https://doi.org/10.1186/s13063-015-0727-9
Jones, B., & Kenward, M. (2003). Design and analysis of cross-over trials. Chapman & Hall/CRC.
Julious SA, McIntyre NE (2012) Sample sizes for trials involving multiple correlated must-win comparisons. Pharm Stat 11:177–185
Koch, M., Riss, P., Umek, W., & Hanzal, E. (2015). The primary outcomes and power calculations in clinical RCTs in urogynecology - need for improvement? *Trials*, *16*(S1). https://doi.org/10.1186/1745-6215-16-s1-p22
Kordzakhia G, Siddiqui O, Huque MF (2010) Method of balanced adjustment in testing co-primary endpoints. Stat Med 29:2055–2066
Lachin, J. (1981). Introduction to sample size determination and power analysis for clinical trials. *Controlled Clinical Trials*, *2*(2), 93-113. https://doi.org/10.1016/0197-2456(81)90001-5
Lu, K., Luo, X., & Chen, P. (2008). Sample Size Estimation for Repeated Measures Analysis in Randomized Clinical Trials with Missing Data. *The International Journal of Biostatistics*, *4*(1).
Makuch R, Simon R. Sample size requirements for evaluating a conservative therapy. Cancer Treat Rep 1978; 62:1037–1040.
Ng, T. (1995). Conventional null hypothesis testing in active control equivalence studies. *Controlled Clinical Trials*, *16*(5), 356-358. https://doi.org/10.1016/0197-2456(95)00040-2
Offen W, Chuang-Stein C, Dmitrienko A, Littman G, Maca J, Meyerson L, Muirhead R, Stryszak P, Boddy A, Chen K, Copley-Merriman K, Dere W, Givens S, Hall D, Henry D, Jackson JD, Krishen A, Liu T, Ryder S,

Sankoh AJ, Wang J, Yeh CH. 2007. Multiple co-primary endpoints: medical and statistical solutions. Drug Inf J 41:31–46
P, F. (2020). *Approaches to sample size estimation in the design of clinical trials--a review. By A. Donner, Statistics in Medicine, 3, 199-214 (1984) - PubMed - NCBI*. Ncbi.nlm.nih.gov. Retrieved 16 May 2020, from https://www.ncbi.nlm.nih.gov/pubmed/8235182.
Patel HI (1991) Comparison of treatments in a combination therapy trial. J BiopharmStat1:171–183 Patterson, S., & Jones, B. *Bioequivalence and statistics in clinical pharmacology*.
Pezzullo JC. Web Pages that Perform Statistical Calculations. Computer Program 2014
PJB Publications. (1991). Pharmacokinetic and bioequivalence studies.
Proschan, M. (2009). Sample size re-estimation in clinical trials. *Biometrical Journal*, *51*(2), 348-357.
Qu, R., & Zheng, H. (2003). Sample size calculation for bioequivalence studies with high-order crossover designs. *Controlled Clinical Trials*, *24*(4), 436-439. https://doi.org/10.1016/s0197-2456(02)00317-3
Rochon, J. (2005). Book Review: Sample size calculations in clinical research. *Clinical Trials: Journal of The Society for Clinical Trials*, *2*(3), 269-270. https://doi.org/10.1191/1740774505cn092xx
Rosner B, Milton RC. Significance Testing for Correlated Binary Outcome Data. Biometrics 1988; 44:505–512. Rosner, B. (1982). Statistical Methods in Ophthalmology: An Adjustment for the Intraclass Correlation between Eyes. *Biometrics*, *38*(1), 105. https://doi.org/10.2307/2530293.
Schneider, B. (1981). The Role of Hypothesis Testing in Clinical Trials. *Methods of Information in Medicine*, *20*(02), 65-66. https://doi.org/10.1055/s-0038-1635299
Schneider, B. (1981). The Role of Hypothesis Testing in Clinical Trials. *Methods of Information in Medicine*, *20*(02), 65-66. https://doi.org/10.1055/s-0038-1635299
Society for clinical trials. (2006). *Clinical Trials: Journal of the Society for Clinical Trials*, *3*(2), 165-247. Song JX (2009) Sample size for simultaneous testing of rate differences in non-inferiority trials with multiple endpoints. Comput Stat Data Anal 53:1201–1207
Sozu T, Kanou T, Hamada C, Yoshimura I (2006) Power and sample size calculations in clinical trials with multiple primary variables. Japan J Biometrics 27:83–96
Sozu T, Sugimoto T, Hamasaki T (2010) Sample size determination in clinical trials with multiple co-primary binary endpoints. Stat Med 29:2169–2179
Sozu T, Sugimoto T, Hamasaki T (2011) Sample size determination in superiority clinical trials with multiple co- primary correlated endpoints. J Biopharm Stat 21:650–668
Sozu T, Sugimoto T, Hamasaki T (2012) Sample size determination in clinical trials with multiple co-primary end points including mixed continuous and binary variables. Biometrical J 5 Springer Nature. 2020. *Clinical Trials*.
Trachtman, H., & Caplan, A. (2018). Data monitoring committees and stopping trials—Giving participants a voice. *Contemporary Clinical Trials*, *68*, 146. https://doi.org/10.1016/j.cct.2018.03.009
Urokinase Pulmonary Embolism Trial Study Group: Urokinase-streptokinase embolism trial: Phase 2 results. JAMA 1974; 229:1606–1613.
Warner, D. (1995). Statistical power, sample size, and their reporting in randomized controlled trials. *Journal of Neurosurgical Anesthesiology*, *7*(1), 69. https://doi.org/10.1097/00008506-199501000-00018
Will my clinical study be a success? The Concept of Statistical Power. CROS NT. (2020). Retrieved 16 May 2020, from https://www.crosnt.com/statistical-power-clinical-trial-success/.
Xiong C, Yu K, Gao F, Yan Y, Zhang Z (2005) Power and sample size for clinical trials when efficacy is required in multiple endpoints: application to an Alzheimer's treatment trial. Clin Trials 2:387–393

****