SPECIAL SECTION

SAMPLES IN RANDOMIZED CLINICAL TRIALS WITH INTERIM ANALYSIS

Michelle Saaibi Meléndez[1,2,a], Felipe Botero-Rodríguez[1,2,a], Carlos Javier Rincón Rodríguez[1,2,b]

[1] Semillero de Bioestadística, School of medicine, Pontificia Universidad Javeriana, Bogotá, Colombia.

[2] Department of Clinical Epidemiology and Biostatistics, Pontificia Universidad Javeriana, Bogotá, Colombia.

[a] Physician; [b] statistician, master in Clinical Epidemiology.

ABSTRACT

This article introduces randomized clinical trials and basic concepts of statistical inference. We present methods for calculating the sample size by outcome type and the hypothesis to be tested, together with the code in the R programming language. We describe four methods for adjusting the original sample size for interim analyses. We sought to introduce these topics in a simple and concrete way, considering the mathematical expressions that support the results and their implementation in available statistical programs; therefore, bringing health students closer to statistics and the use of statistical programs, which are aspects that are rarely considered during their training.

Keywords: Sample Size; Clinical Trials; Hypothesis Tests (source: MeSH NLM).

INTRODUCTION

The approach to medicine has shifted from an initial paternalistic view to pragmatic reductionism. This change occurred because of the drive to improve the quality of care, decrease individual economic incentives and prioritize the importance of research to improve the quality of evidence [1]. Evidence-based medicine emerged as a new paradigm in the 1990s as scientific support for clinical decision-making and is based on a hierarchy of three statements: a) randomized clinical trials (RCTs) or systematic reviews of many experiments usually provide more evidence than observational studies; b) analytical clinical studies provide better evidence than pathophysiological rationale alone; and c) analytical clinical studies provide more evidence than expert judgment [2].

Obtaining valid results from RCTs depends on the quality of the data, which must be sufficient to address the research question. To obtain these quality data, the sample size must be large enough to obtain an accurate estimate of the effect of the intervention. Random errors will not affect the interpretability of the results as long as the sample is large enough; however, a systematic error can invalidate a study [3].

The interim analysis consists of setting an observation point(s), so the behavior of the sample can be assessed up until that point. Depending on the results, the committee may determine if the study is relevant enough to continue or not [3]. This article seeks to provide an introduction to the calculation of sample size by type of outcome and hypothesis. We also aim to provide information on its adjustment by interim analysis, considering the mathematical

Cite as: Saaibi Meléndez M, Botero-Rodríguez F, Rincón Rodríguez CJ. Samples in randomized clinical trials with interim analysis. Rev Peru Med Exp Salud Publica. 2023;40(2):220-8. doi: 10.17843/rpmesp.2023.402.12217.

Correspondence:
Michelle Saaibi Meléndez;
msaaibi@javeriana.edu.co

Received:11/10/2022
Approved: 26/04/2023
Online: 30/06/2023

https://doi.org/10.17843/rpmesp.2023.402.12217.

formulas and their implementation in available statistical programs such as the R programming language. The objective is to bring health personnel closer to statistics and the use of programs, aspects that are little considered in their training. Although there are already several sources that develop the above topics, there are not many documents that merge both theory and practice, including all the aspects mentioned above regarding RCTs. Reviewing articles allows young researchers and health professionals to make a first approach to these topics, without generating an initial rejection due to their complexity. Connecting mathematical expressions with their implementation in a statistical program seeks to avoid that, once again, young researchers execute pre-established functions such as TwoSampleMean.Equality or TwoSampleMean.NIS (included in the "TrialSize" package [4]) without understanding where the results come from, the effect of the parameters on the sample size or the need to choose parameters with values consistent with the type of hypothesis that is being evaluated. This seeks to promote understanding of the mechanical execution of tasks only to meet the requirements of an evaluation committee.

Randomized clinical trials

The equipoise principle corresponds to a state of uncertainty regarding the therapeutic results of a treatment, which justifies an RCT [5]. RCTs with a control group are prospective studies that compare the outcomes of an intervention(s) with the best available alternative. In these studies, patient safety should always be a priority, so the possible benefits, harms and treatment alternatives for the patient's condition should be explained. Although it may have limitations, it is considered the best alternative for evaluating the efficacy or safety of an intervention [6,7]. It is characterized by: a) an intervention that is compared with a control group that can be placebo or the usual treatment, b) randomized assignment of the interventions in the population to reduce possible confusion bias by obtaining homogeneous groups and the possibility of selection bias by avoiding foreseeing the group to which the patient is assigned c) the blinding of the treatment groups can be performed both for researchers, patients or analysts, which minimizes possible information biases [6,7].

The RCTs are divided into four phases. Phase I seeks to determine possible toxic effects, absorption, distribution and metabolism of the drug in a group of 20-80 healthy people. Phase II is conducted in a diseased population to determine the safety and efficacy of the drug, based on biological markers and evaluating adverse reactions. Phase III is performed when there is evidence on the safety and efficacy of the intervention and additional information is sought on the safety and effectiveness of the drug in a larger number of participants. The intervention is compared with the usual therapy or placebo in a long-term follow-up in order to identify possible side effects. In phase IV, after the molecule has been approved for marketing, it is compared with other existing products in the general population; pharmacovigilance is also carried out in order to look for adverse events not identified in phase III due to their low incidence or long periods of occurrence [3,6].

In this article we will focus on phase III and IV studies, which require a sample size calculation. Additionally, we will work on parallel RCTs characterized by a simultaneous follow-up of each group to which they were assigned [3].

Inferential statistics

Inferential statistics allows estimating the behavior of the entire population from the results obtained in a sample. This behavior is summarized in measures such as means, proportions or variances, which, if obtained for the whole population, would be called parameters [7]. There are two alternatives: confidence intervals and hypothesis tests; with consistent results, the first seeks a range of values that, with a degree of confidence, contains the parameter of interest, while the second evaluates a statement about the parameter of interest, making the decision to reject it or not.

Since this paper presents the sample size calculation in parallel RCTs to evaluate parameter statements, we will describe the process of hypothesis testing. Initially, two hypotheses are proposed, the null hypothesis (H_0) which is a statement about the parameter, and the alternative hypothesis (H_a) which is its negation; almost always the alternative hypothesis [8], which is related to the research question is sought to be tested; at the end a decision is made to reject or not the H_0. Taking into account that this decision depends on the results obtained only from a sample, there is the probability of committing two errors, type I error or significance level (α) that occurs when rejecting H_0 when it is true, and type II error (β), occurs when not rejecting H_0 when it is false [9]. The opposite of type I error is the confidence level ($1-\alpha$) and corresponds to the probability of not rejecting H_0 when it is true, and the opposite of type II error ($1-\beta$), which is the power, is the probability of rejecting H_0 when it is false [9]. When performing a hypothesis test, the probability of committing the type I and II error is low, which implies that the confidence level and power have a high probability (typically: $\alpha=0.05$ and

β=0.1 or 0.2). In order to guarantee these values, it is necessary to calculate the sample size.

In order to make the decision to reject H_0, an operation is carried with the values from the sample (test statistic) and contrasted with the behavior that should occur if H_0 were true. If the value found by the test statistic is unlikely, this is evidence that H_0 is false and is rejected in favor of H_a, otherwise it is considered that there is not enough evidence to reject H_0. The probability reflecting the evidence "for" or "against" H_0 is called the p-value [10,11], and is equal to the probability, assuming the null hypothesis is true, of obtaining a value of the test statistic "...as extreme or more (in the appropriate direction of H_a) than the value actually calculated" [11,12]; finally, H_0 is rejected under the assumption of a value of $p<\alpha$. Statistical significance, commonly evaluated by means of the p-value, does not account for clinical significance; we speak of statistical significance when the premise of a value of $p<\alpha$ is fulfilled, while clinical significance is defined by those results that improve the physical, mental and social functionality of the patient, which can lead to an improvement in the quality or quantity of life, depending on the context [14].

Types of hypotheses

There can be different research questions in an RCT that relate to four different ways of stating the H_0. The sample size calculation depends on the type of hypothesis to be tested; therefore, Table 1 presents their definitions along with an example.

Sample size

Generally, it is not possible to study the entire population, therefore, a specific sample size (n) is required to represent its behavior. As the sample size increases, the results approach that of the population, so that from a specific size, the results will not present large changes, making it unnecessary to continue collecting participants [15]. Recruiting more subjects than necessary increases both the complexity of the logistical operation and the costs, and poses an ethical dilemma by unnecessarily assigning subjects to a treatment that has not proven its benefit. On the other hand, defining a very small sample size implies a high risk of the type II error mentioned above. The calculation of the sample size makes it possible to determine whether a study is feasible based on *a priori* assumptions, given the power, significance and background of previous studies addressing the same research question, taking into account the ethical considerations of subjecting people to an experiment [13,16].

In addition, when conducting an RCT, the possibility arises of observing the results obtained as the sample is collected. This is called "interim analyses", which should be planned from the beginning of the investigation during the preparation of the protocol. These additional analyses increase the possibility of type I and II errors, and for this reason, the sample size must be adjusted to maintain a level of confidence and overall power throughout the RCT. The above reflects the importance of the sample size calculation, therefore, this article presents how to calculate the sample size for RCTs, showing the expression from which it is obtained, and its application using the R programming language [17]. Additionally, we present how to perform the adjustment for interim analysis together with an example.

MATERIALS AND METHODS

Based on the review of the book "Sample size calculations in clinical research" by Chow *et al.* [13], this article presents how to calculate the sample size for a parallel RCT, by: 1) type of outcome (dichotomous, continuous) and 2) type of hypothesis to be evaluated (equality, non-inferiority, superiority and equivalence). The corresponding mathematical expressions and the code to create a function in the R [17] and RStudio [18] programs are included. For the use of this code, the reader is required to have a basic knowledge of the use of these programs, where the function must be copied and executed; the function can then be used including the required parameters described in the results section. For each scenario, an example with fictitious data is included, specific considerations related to the function parameters are mentioned as well.

The methods of Pocock, O'Brien and Fleming, Wang and Tsiatis and Inner Wedge to adjust the original sample size obtained from the functions created previously in order to perform the interim analysis are described below. The adjustment consists of multiplying the original sample size by the coefficients included in Annexes 1 to 4 depending on the method used, and considering the number of planned evaluations (R), the power and significance level defined for the study. In addition, the expression of the test statistic used for each evaluation by type of outcome is included, based on the information of the participants entering the study. In summary, we present the following for each method: 1) the critical values that correspond to the values of the standard normal distribution that determine the rejection zone for evaluating the null hypothesis at each point in time, and 2) the coefficients for adjusting the sample size calculation.

https://doi.org/10.17843/rpmesp.2023.402.12217.

Table 1. Types of hypotheses in randomized clinical trials.

Type	Definition [24]	Hypothesis [25,26]	Example	
			Hypothesis	Interpretation
Equality	Evaluates whether there are differences between the treatment and control groups.	H_0: There is no difference between the two therapies.	Pressure over the estimated sternal projection of the aortic valve at the sternum is not associated with a change in hemodynamic parameters in the hypotensive patient.	Patients who underwent a pressure of 6 mm depth over the estimated sternal projection of the aortic valve on the sternum, maintained for 90 seconds, showed a homogeneous decrease of blood pressure and heart rate parameters [27].
		H_a: There is a difference between the two therapies.	Pressure over the estimated sternal projection of the aortic valve on the sternum is associated with a change in hemodynamic parameters in the hypotensive patient.	
Non-inferiority	It evaluates whether the effect of a new treatment (whose effect is lower than the conventional treatment, but greater than the placebo) is within an accepted range and is established on the basis of the best available evidence. This difference is justified by side effects or feasibility.	H_0: The effect of the new intervention is less than or equal to the placebo.	The new antimicrobial has the same effectiveness as the placebo.	The new antimicrobial, although better tolerated than conventional therapy, is less effective clinically and statistically, so it cannot be recommended as first line [28].
		H_a: The effect of the new intervention is greater than the placebo.	The new antimicrobial is more effective than the placebo.	
Superiority	Seeks to evaluate whether a new intervention generates better clinical outcomes than a well-established therapy or placebo.	H_0: The new intervention is not superior to the established therapy.	Volunteering does not reduce social isolation or impact better mental health outcomes.	Volunteering did not prove to be superior compared to the control group regarding mental health outcomes or isolation [29].
		H_a: New intervention is superior to established therapy.	Volunteering reduces social isolation and impacts better mental health outcomes.	
Equivalence	It seeks to evaluate whether the effect of the treatment is identical to that of another therapy.	H_0: Therapies are not equivalent.	The inclusion of metformin, associated with oral contraceptives in the treatment of polycystic ovary syndrome, is not as effective as monotherapy with oral contraceptives alone.	The ultrasound remission time was shorter, there were less symptoms and the recurrence rate at 3 months was lower with the combined therapy, which shows greater effectiveness compared to the study group that received monotherapy [30].
		H_a: The therapies are equivalent.	Oral contraceptive monotherapy is as effective as oral contraceptive therapy plus metformin for the treatment of polycystic ovary syndrome.	

H_0: null hypothesis, H_a: alternative hypothesis

Sample size calculation for a dichotomous outcome

As an example, we assume that two treatments are to be compared and the outcome of interest is the proportion of deaths. For all expressions below, we denote p_T and p_C as the proportion of deceased in the treatment and control group, respectively; ϵ is the expected difference between these two proportions $(\epsilon=p_T-p_C)$, δ is the margin of tolerance or superiority defined by the researchers, and k is the ratio between the sample size of the treatment group and the control group $(k=n_T/n_C)$, i.e., $n_T=kn_C$. Finally, we denote α and β as the type I and II error, respectively; and $z_{(q)}$ as the q percentile of the standard normal distribution function. In Table 2, we present the expressions to obtain n_C, and, in the inset, the code in the R programming language that creates a function for its implementation, along with an example where $\alpha=0.05$, $\beta=0.2$ and $k=1$.

In all four hypotheses, the smaller the expected difference (ϵ) and the closer the proportions are to 0.5, the larger the sample size. When testing a non-inferiority hypothesis, if the higher the proportion of the event the greater the effectiveness, then $\delta<0$; if the lower the proportion of the event the greater the effectiveness, then $\delta>0$. When testing a superiority hypothesis, if the higher the proportion of the event the greater the effectiveness, then $\delta>0$; if the lower the proportion of the outcome the greater the effectiveness, then $\delta<0$. When testing an equivalence hypothesis, always $\delta>0$.

Continuous outcome sample size calculation

As an example, we assume that two treatments are to be compared, and the outcome is systolic blood pressure in mmHg (SBP). For all the expressions presented below, we denote μ_T and μ_C as the mean SBP in the treatment and control group, respectively; ϵ is the expected difference between the two means $\epsilon=\mu_T-\mu_C$ and is the standard deviation of the two samples together. δ, k, α, β and $z_{(q)}$ represent the same values as in the previous section. In Table 3, we present the expressions to obtain the code in the R programming language for implementation with an example where $\alpha=0.05$, $\beta=0.2$ and $k=1$.

In all four hypotheses, higher s and lower ϵ require larger sample sizes. While testing a hypothesis of noninferiority, if the higher μ the greater the effectiveness, then $\delta<0$; if the lower μ the greater the effectiveness, then $\delta>0$. When testing a superiority hypothesis, if the higher μ the greater the effectiveness, then $\delta>0$; if the lower μ the greater the effectiveness, then $\delta<0$. When testing an equivalence hypothesis, always $\delta>0$.

RESULTS

Interim Analysis

In an RCT, the study hypothesis can be tested sequentially as the sample is collected, giving the possibility of interrupting the collection if a clear benefit of the intervention is identified early. Depending on the number of evaluations (R) that are programmed, it is necessary to adjust the initial sample size to maintain the overall significance level of the study, and to establish the critical values on the distribution of the test statistic to reject or accept the null hypothesis in each evaluation. R evaluations are performed as $\frac{n}{R}$ subjects accumulate, and the test statistic z_r $(r=1,2,...,R)$ for a dichotomous outcome is equal to:

$$z_r = \frac{\sqrt{n_r}(\hat{p}_{T,r}-\hat{p}_{C,r})}{\sqrt{\hat{p}_{T,r}(1-\hat{p}_{T,r}) + \hat{p}_{C,r}(1-\hat{p}_{C,r})}}$$

where n_r, $\hat{p}_{T,r}$ and $\hat{p}_{C,r}$ are the sample size per intervention group and the estimated proportions of the outcome at the r time of assessment of the treatment group and the control group, respectively.

For a continuous outcome, the test statistic is equal to:

$$z_r = \frac{1}{\sqrt{n_r(\hat{s}^2_{Tr} + \hat{s}^2_{Cr})}}\left(\sum_{j=1}^{n_r} x_{Tj} \sum_{j=1}^{n_r} x_{cj}\right)$$

where n_r, $\hat{s}^2_{Tr}$ and $\hat{s}^2_{Cr}$ are the sample size in each intervention group, and the estimated variances at the time of the r_th evaluation of the treatment group and the control group, respectively. x_{Tj} and x_{Cj} are the observed values of the outcome in each subject collected until time r.

We present four methods that allow the adjustment of the sample size depending on the number of programmed evaluations, the significance level and the power established in a hypothesis of equality. First, the Pocock method, in which the sample size is adjusted by multiplying the sample size initially obtained from the expressions presented in the previous section, by the coefficients included in Annex 1, depending on the number of evaluations and the significance and power levels established. Now, if $|z_r|>CP_{(r,\alpha)}$, the H_0 is rejected and data collection is suspended, otherwise, the collection continues. The critical values $CP_{(r,\alpha)}$ are presented in Annex 1 for defined R and α. The second method is that of O'Brien and Fleming and the coefficients to perform the initial sample size adjustment are presented in Annex 2. In this approach H_0 is rejected at each evaluation if $|z_r|>COF_{(r,\alpha)}$

https://doi.org/10.17843/rpmesp.2023.402.12217.

Table 2. Types of hypotheses with dichotomous outcome and their code in the R programming language.

Types of hypotheses	Code in R programming language

Equality hypothesis
$n_c = \frac{\left(z_{(1-\frac{\alpha}{2})} + z_{(1-\beta)}\right)^2}{\epsilon^2}\left[\frac{p_T(1-p_T)}{k} + p_C(1-p_C)\right]$
Then, if it is expected that the proportion of deaths in the treatment group and in the control group are equal to $p_T = 0.15$ and $p_C = 0.2$, $n_C=n_T=903$.
	n.2prop.igual<-function(alpha,beta,k,pT,pC){
	nC<-(qnorm(1-alpha/2)+qnorm(1-beta))^2/(pT-pC)^2*(pT*(1-pT)/k+pC*(1-pC))
	nT<-k*nC
	
	Grupo<-c("Tratamiento=","Control=")
	n<-ceiling(c(nT,nC))
	n<-data.frame(Grupo,n)
	print(n)
	}

Non-inferiority hypothesis
$n_c = \frac{(z_{(1-\alpha)} + z_{(1-\beta)})^2}{(\epsilon-\delta)^2}\left[\frac{p_T(1-p_T)}{k} + p_C(1-p_C)\right]$
Then, if the expected proportion of deaths in the treatment group and in the control group are p_T=0.2 and p_C=0.22 and an increase in mortality is tolerated from δ=0.03, $n_C=n_T$=821.
	n.2prop.noinf<-function(alpha,beta,k,pT,pC,delta){
	nC<-(qnorm(1-alpha)+qnorm(1-beta))^2/((pT-pC)-delta)^2*(pT*(1-pT)/k+pC*(1-pC))
	nT<-k*nC
	
	Grupo<-c("Tratamiento=","Control=")
	n<-ceiling(c(nT,nC))
	n<-data.frame(Grupo,n)
	print(n)
	}

Superiority hypothesis
$n_c = \frac{(z_{(1-\alpha)} + z_{(1-\beta)})^2}{(\epsilon-\delta)^2}\left[\frac{p_T(1-p_T)}{k} + p_C(1-p_C)\right]$
Then, if it is expected that the proportion of deaths in the treatment group and in the control group are p_T=0.18 and p_C=0.25, and it is considered superior if it reduces mortality by at least δ=- 0.01, $n_C=n_T$=576.
	n.2prop.sup<-function(alpha,beta,k,pT,pC,delta){
	nC<-(qnorm(1-alpha)+qnorm(1-beta))^2/((pT-pC)-delta)^2*(pT*(1-pT)/k+pC*(1-pC))
	nT<-k*nC
	
	Grupo<-c("Tratamiento=","Control=")
	n<-ceiling(c(nT,nC))
	n<-data.frame(Grupo,n)
	print(n)
	}

Equivalence hypothesis
$n_c = \frac{(z_{(1-\alpha)} + z_{(1-\beta/2)})^2}{(\delta - \lvert\epsilon\rvert)^2}\left[\frac{p_T(1-p_T)}{K} + p_C(1-p_C)\right]$
Then, if it is expected that the proportion of deaths in the treatment group and in the control group are p_T=0.22 and p_C=0.18 and they are defined as equivalent if they do not differ by more than $\lvert\delta\rvert$=0.1, $n_C=n_T$=760.
	n.2prop.equi<-function(alpha,beta,k,pT,pC,delta){
	nC<-(qnorm(1-alpha)+qnorm(1-beta/2))^2/(delta-abs(pT-pC))^2*(pT*(1-pT)/k+pC*(1-pC))
	nT<-k*nC
	
	Grupo<-c("Tratamiento=","Control=")
	n<-ceiling(c(nT,nC))
	n<-data.frame(Grupo,n)
	print(n)
	}

$\sqrt{(R/r)}$, otherwise it continues. The critical values $COF_{(r,\alpha)}$ are presented in Annex 2 according to the number of evaluations and significance level. The third method is that of Wang and Tsiatis, which includes a new parameter Δ ; coefficients for sample size adjustment for α=0.05 are included in Appendix 3. In this method H_0 is rejected if $|z_r| > CWT_{(r,\alpha,\Delta)}\left(\frac{r}{R}\right)^{\Delta-0,5}$ otherwise it continues. The critical $CWT_{(r,\alpha,\Delta)}$ values are presented in Annex 3 for α=0.05. The methods of Pocock and O'Brien and Fleming are particular cases of the method of Wang and Tsiatis when Δ=0.5 and Δ=0, respectively, therefore, the critical values for these values of Δ are obtained from Annexes 1 and 2.

Finally, we present the Inner Wedge method; in this method unlike the previous three, two critical values are proposed: if $|zr| \geq b_r$ rejects H_0 and collection is suspended, the conclusion is that a significant treatment effect was found, on the other side, if $|zr| < a_r$ does not reject H_0 and collection is suspended the conclusion is that no differences between treatment and control are going to be found, otherwise, collection continues. The critical values a_r and b_r are equal to:

$a_r=[Cw1_{(r,\alpha,\beta,\Delta)} + Cw2_{(r,\alpha,\beta,\Delta)}]\sqrt{\frac{r}{R}} - Cw2_{(r,\alpha,\beta,\Delta)}\left(\frac{r}{R}\right)^{(\Delta-0,5)}$, if $a_r<0 \Rightarrow a_r=0$

and $b_r=Cw1_{(r,\alpha,\beta,\Delta)}\left(\frac{r}{R}\right)^{(\Delta-0,5)}$

Table 3. Types of hypotheses by continuous outcome and code in R programming language..

Type of hypothesis	Code in R programming language

Equality hypothesis
$n_C=\frac{(z_{(1-\alpha/2)}+z_{(1-\beta)})^2\, s^2\, (1+1/k)}{\epsilon^2}$
Then, if the mean μ_T=150 and μ_C=160 and s=28, $n_C=n_T$=124.
	n.2mu.igual
	-function(alpha,beta,k,muT,muC,s){
	nC<-(qnorm(1-alpha/2)+qnorm(1-
	beta))^2*s^2*(1+1/k)/(muT-muC)^2
	nT<-k*nC
	
	Grupo<-c("Tratamiento =","Control =")
	n<-ceiling(c(nT,nC))
	n<-data.frame(Grupo,n)
	print(n)
	}

Non-inferiority hypothesis
$n_C=\frac{(z_{(1-\alpha)}+z_{(1-\beta)})^2\, s^2\, (1+1/k)}{(\epsilon-\delta)^2}$
Then, if μ_T=155, μ_C =160 and s=28, and is defined to be non-inferior if the maximum increases by δ=5, $n_C=n_T$=97.
	n.2mu.noinf<-
	function(alpha,beta,k,muT,muC,s,delta){
	nC<-(qnorm(1-alpha)+qnorm(1-
	beta))^2*s^2*(1+1/k)/((muT-muC)-delta)^2
	nT<-k*nC
	
	Grupo<-c("Tratamiento =","Control =")
	n<-ceiling(c(nT,nC))
	n<-data.frame(Grupo,n)
	print(n)
	}

Superiority hypothesis
$n_C=\frac{(z_{(1-\alpha)}+z_{(1-\beta)})^2\, s^2\, (1+1/k)}{(\epsilon-\delta)^2}$
Then, if μ_T=145, μ_C =160 and s=28, and is considered superior if it at least decreases by δ=-10, $n_C=n_T$= 388.
	n.2mu.sup<-
	function(alpha,beta,k,muT,muC,s,delta){
	nC<-(qnorm(1-alpha)+qnorm(1-
	beta))^2*s^2*(1+1/k)/((muT-muC)-delta)^2
	nT<-k*nC
	
	Grupo<-c("Tratamiento =","Control =")
	n<-ceiling(c(nT,nC))
	n<-data.frame(Grupo,n)
	print(n)
	}

Equivalence hypothesis
$n_C=\frac{(z_{(1-\alpha)}+z_{(1-\beta/2)})^2\, s^2\, (1+1/k)}{(\delta-\lvert\epsilon\rvert)^2}$
Then, if μ_T=150, μ_C =160 and s=28, and they are defined as equivalent if they do not differ by more than de $\lvert\delta\rvert$=5, $n_C=n_T$=538.
	n.2mu.equi<-function(alpha,beta,k,muT,muC,s,delta){
	nC<-(qnorm(1-alpha)+qnorm(1-beta/2))^2*s^2*(1+1/k)/(delta-abs(muT-
	muC))^2
	nT<-k*nC
	
	Grupo<-c("Tratamiento =","Control =")
	n<-ceiling(c(nT,nC))
	n<-data.frame(Grupo,n)
	print(n)
	}

Annex 4 presents the values of *Cw*1 and *Cw*2 for an α=0.05 and a power of 0.8 and 0.9, and the columns *coef.fit* include the coefficients, by which the original sample must be multiplied to perform the R evaluations.

As an example, consider that you want to compare drug A vs. placebo and the outcome is the proportion of deaths at the end of follow-up. In a hypothesis of equality, assuming $\alpha=0.05$, $\beta=0.1$, $p_T=0.1$ and $p_C=0.2$ (i.e. ϵ=0.1), for two groups of the same size, k=1, the required sample size in each group is 263 subjects. If we plan to perform R=5 evaluations, the adjusted sample size by Pocock's method is 263 x 1.207=318 for each group and the critical value in each evaluation is $CP_{0.05,r}$=2.413. The adjusted sample size by the method of O'Brien and Fleming is 263 x 1,026=270 for each group and the critical values for each evaluation are $COF_{(r,0,05)}$=4.562; 3.226; 2.634; 2.281 and 2.040. The adjusted sample size with the method of Wang and Tsiatis for each group, with Δ=0.25, would be 263 x 1.066=281, and the critical values at each evaluation would be $CWT_{(r;0.05;0.25)}$=3.194; 2.686; 2.427; 2.259 and 2.136. Finally, the adjusted sample size with the Inner Wedge method for each group would be, with Δ=0.25, 263 x 1.199=316, and the critical values for each evaluation are a_r=0;0.388; 1.072; 1.613 and 2.073 and b_r=3.1; 2.607; 2.355; 2.192 and 2.073.

DISCUSSION

In this article we present an approach to sample size adjustment by interim analysis in parallel RCTs, starting from the calculation of the original sample size for subsequent adjustment by one of the four methods described. This paper is aimed at students and young researchers, mainly from the health area, who will find in this article an initial context on RCTs and a review of the main concepts of statistical inference from hypothesis testing. We seek, in a simple and concrete way, to provide an introduction to this topic, integrating the different aspects such as the mathematical expressions that support the results and their implementation in availa-

https://doi.org/10.17843/rpmesp.2023.402.12217.

ble statistical programs. Although there are other resources available for the calculation of sample size such as Internet pages [19] or packages in the R programming language [4], mainly in languages other than Spanish, we found that providing the possibility of using statistical programs that allow students to apply the theory gives a greater understanding of these topics, as opposed to following a sequence of steps in a mechanical way, often without understanding what is generated by the different programs or resources available. This brings students of health areas closer to statistics and the use of statistical programs, an aspect often considered not important during their training.

This article allows the reader to plan an RCT in parallel by defining the sample size and allowing the results to be monitored during the course of the study. At this point, we recommend reviewing additional methods that provide more flexibility, for example, planning the intermediate evaluations on specific dates and not when a fixed number of participants in both groups are completed, which is the main restriction for the four methods presented in this article. The method proposed by Lan and DeMets [20] and R programming language packages such as gsDesign [21] would be interesting material to further explore these issues.

Finally, the decision to discontinue the execution of an RCT, either because of great benefits, potential harms or if it is very unlikely to obtain benefits (futility), should be taken by a group of people independent of the researchers, made up of experts in the clinical area under study, in methodological aspects such as epidemiologists or biostatisticians, and in ethical aspects [3,22,23]. All the necessary aspects should be considered in this decision and not only the result of the evaluation of a statistical test. Planning an interim analysis by adjusting the sample size when the study is being designed will allow supporting, to a greater extent, the value of this criterion during the decision-making process.

Authorship contributions. All authors declare that they meet the authorship criteria recommended by the ICMJE.

Roles according to CRediT. MSM: Conceptualization. Methodology. Investigation. Writing – original draft. Writing – review and editing. Visualization. FBR: Methodology. Investigation. Writing – original draft. Visualization. CJR: Conceptualization. Methodology. Software. Investigation. Writing – original draft. Writing – review and editing.

Funding. Self-funding.

Conflicts of interest. The authors declare that they have no conflicts of interest.

Supplementary material. Available in the digital version of the RPMESP.

REFERENCES

1. F Berwick DM. Era 3 for Medicine and Health Care. Obstet Gynecol Surv. 2016;315(13):1329–30. doi:10.1001/jama.2016.1509.
2. Gordon G, Drummond R, Maureen OM, Deborah JC. Users' Guides to the Medical Literature, 3rd ed [Internet]. McGraw Hill; 2008 [cited 2022 Oct 11]. Available from: https://jamaevidence.mhmedical.com/content.aspx?bookid=847&sectionid=69030714.
3. Friedman LM, Furberg CD, DeMets DL, Reboussin DM, Granger CB. Fundamentals of clinical trials. In: Fundamentals of Clinical Trials [Internet]. Spinger; 2015 [cited 2022 Oct 11]. Available from: https://link.springer.com/book/10.1007/978-3-319-18539-2.
4. Zhang E, Wu VQ, Chow S-C, Zhang HG. TrialSize: R Functions for Chapter 3,4,6,7,9,10,11,12,14,15 of Sample Size Calculation in Clinical Research_. R package version 1.4 [Internet]. 2020 [cited 2022 Oct 11]. Available from: https://cran.r-project.org/package=TrialSize.
5. Miller FG, Joffe S. Equipoise and the Dilemma of Randomized Clinical Trials. N Engl J Med. 2011;364(5):476–80. doi: 10.1056/nejmsb1011301.
6. Lazcano-Ponce E, Salazar-Martínez E, Gutiérrez-Castrellón P, Angeles-Llerenas A, Hernández-Garduño A, Viramontes JL. Ensayos clínicos aleatorizados: variantes, métodos de aleatorización, análisis, consideraciones éticas y regulación. Salud Publica Mex. 2004;46(6):559–84.
7. Diana MF. Conceptos básicos sobre bioestadística descriptiva y bioestadística inferencial. Rev Argentina Anestesiol. 2006;64(6):241–51.
8. Flight L, Julious SA. Practical guide to sample size calculations: An introduction. Pharm Stat. 2016;15(1):68–74. doi: 10.1002/pst.1709.
9. Cohen HW. P values: Use and misuse in medical literature. Am J Hypertens. 2011;24(1):18–23. doi: 10.1038/ajh.2010.205.
10. Prel J-B du, Hommel G, Röhrig B, Blettner M. Confidence interval or P value. Dtsch Arztebl. 2009;106(19):335–9. doi: 10.3238/arztebl.2009.0335.
11. Goodman S. A Dirty Dozen: Twelve P-Value Misconceptions. Semin Hematol. 2008;45(3):135–40. doi: 10.1053/j.seminhematol.2008.04.003.
12. Wayne DW. Bioestadística. Base para el análisis de las ciencias de la salud. 4ta ed. LIMUSA WILEY. 2006.
13. Chow S-C, Shao J, Wang H, Lokhnygina Y. Sample size calculations in clinical research. 2nd ed. Chapman & Hall; 2008.
14. Armijo-Olivo S, Warren S, Fuentes J, Magee DJ. Clinical relevance vs. statistical significance: Using neck outcomes in patients with temporomandibular disorders as an example. Man Ther. 2011;16:563–72. doi: 10.1016/j.math.2011.05.006.
15. Chan A, Tetzlaff JM, Altman DG, Laupacis A, Gøtzsche PC, Krleža-jeric K, et al. Declaración SPIRIT 2013: definición de los elementos estándares del protocolo de un ensayo clínico. Rev Panam salud pública. 2015;38(6):506–14.
16. Gowda GS, Komal S, Sanjay TN, Mishra S, Kumar CN, Math SB. Sociodemographic, legal, and clinical profiles of female forensic inpatients in Karnataka: A retrospective study. Indian J Psychol Med. 2019;41(2):138–43. doi: 10.4103/IJPSYM.IJPSYM_152_18.
17. R Core Team. R: A language and environment for statistical computing [Internet]. R Foundation for Statistical Computing, Vienna, Austria; 2022 [cited 2022 Oct 11]. Available from: https://www.r-project.org/.
18. RStudioTeam. RStudio: Integrated Development Environment for R [Internet]. RStudio, PBC, Boston, MA; 2022 [cited 2022 Oct 11]. Available from: http://www.rstudio.com/.
19. Sample Size Calculator [Internet]. Cleveland Clinic; 2022 [cited 2022 Oct 11]. Available from: https://riskcalc.org/samplesize/.
20. Demets DL, Lan KKG. Interim analysis: the alpha spending approach. Stat Med. 1994;13:1341–52. doi: 10.1002/sim.4780131308.
21. Anderson K. gsDesign: Group Sequential Design [Internet]. R package version 3.4.0; 2022 [cited 2022 Oct 11]. Available from: https://CRAN.R-project.org/package=gsDesign.

22. Ellenberg S, Fleming T, DeMets D. Data Monitoring Committees in Clinical Trials: A Practical Perspective. Wiley; 2003.
23. Fisher M, Roecker E, DeMets D. The role of an independent statistical analysis center in the industry-modified National Institutes of Health model. Drug Inf J. 2001;(35):115–29. doi: 10.1177/009286150103500113.
24. Kumbhare D, Alavinia M, Furlan J. Hypothesis Testing in Superiority, Noninferiority, and Equivalence Clinical Trials. Am J Phys Med Rehabil. 2019;98(3):226–30. doi: 10.1097/PHM.0000000000001023.
25. Services H. Non-Inferiority Clinical Trials to Establish Effectiveness Guidance for Industry [Internet]. FDA; 2016 [cited 2022 Oct 11]. Available from: https://www.fda.gov/media/78504/download.
26. Flight L, Julious SA. Practical guide to sample size calculations: an introduction. Pharm Stat. 2016;15(1):68–74. doi: 10.1002/pst.1709.
27. Benito MM, Marín RC. Cambios en la presión arterial y frecuencia cardíaca después de una presión sobre la válvula aórtica en sujetos con hipertensión arterial esencial. Osteopat Cient. 2008;3(3):100–7. doi: 10.1016/S1886-9297(08)75758-8.
28. Chadwick D. Safety and efficacy of vigabitrin and carbamazepine in newly diagnosed epilepsy: a multicentre randomised double-blind study. Lancet. 1999;354:13–9. doi: 10.1016/s0140-6736(98)10531-7.
29. Priebe S, Chevalier A, Hamborg T, Golden E, King M, Pistrang N. Effectiveness of a volunteer befriending programme for patients with schizophrenia: Randomised controlled trial. Br J Psychiatry. 2019;1–7. doi: 10.1192/bjp.2019.42.
30. Bascope EL, Ortiz YM, Llanos GRL, Lizbeth MHA, Lazo L. Metformina en el tratamiento del síndrome de ovarios poliquísticos. Un ensayo clínico aleatorizado. Rev Cient Cienc Med. 2017;20(2):45–52.

https://doi.org/10.17843/rpmesp.2023.402.12217.