Sample size in clinical protocols

Alvar Loria

Secretaría de Salud, Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Clinical Epidemiology Unit, Ciudad de México, Mexico

Abstract

Sample size calculation in clinical protocols submitted to the review board of the Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán frequently has three types of problems: insufficient data for sample size recalculation, use of obsolete equations that do not take the beta error into account, and use of equations that do not correspond to the study design. In this document, we discuss the four data that should be included in the clinical protocols of randomized controlled trials that compare drugs versus placebo in superiority trials, and versus active controls in non-inferiority studies.

KEY WORDS: Sample size. Ethics. Clinical research. Non-inferiority trials.

Knowledge rests not upon truth alone, but upon error also.

Carl Gustav Jung

Why is sample size (SS) calculation important?

Currently, including a deficit or an excess of study subjects in clinical investigation is considered to be an ethical fault. A number must be included that allows to detect, with relative certainty, the effect of a medical intervention if there is any, and simultaneously, relative certainty that if there is no effect, the intervention is truly negative.

Excess of study subjects increases the odds of iatrogenic harm to the subjects and, in addition, it makes studies more expensive. Its opposite, an insufficient sample, with low power, can drive to negative clinical studies, which leads to lower probability of being published and this, in turn, to the so-called publication bias that translates into a deficit of negative studies.¹

Publication bias leads to falsely increasing the effectiveness of any therapeutic intervention, which drove collegiate bodies to make repeated appeals to researchers and editors encouraging them to publish

the results of studies where a medication is not different from placebo.

First statement

An inadequate sample size has important repercussions, not only ethical and economic, but also in clinical trials interpretation and conclusions.

In the annual statistics of protocols submitted to research and ethics committees of *Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán*, the SS calculation has always occupied the first place in flaws. The most important are insufficient information to recalculate the SS (the most common), use of obsolete equations that do not take the beta error into account and lack of correspondence between the equation and the study design (the least common).

First world researchers are not safe from problems as neither are protocols supported by the pharmaceutical industry; for example: in 2009, only a minority, 188 of 446 (42%) of the protocols submitted to ethics committees of the United Kingdom provided full information that allowed to recalculate the SS and there was no big difference between industry-supported and non-commercial protocols (44 versus 40 %,

Correspondence:

Alvar Loria

E-mail: alvarloria@gmail.com

Date of reception: 04-08-2017

Date of acceptance: 13-02-2018

DOI://dx.doi.org/10.24875/GMM.M18000163

Gac Med Mex. 2018;154:345-348

Contents available at PubMed

www.gacetamedicademexico.com

respectively). On the other hand, having complete data was no guarantee, since in 54 of the 188 complete ones (29 %) a different SS was obtained when recalculating it.²

Second statement

The most common flaw of protocols is that they do not provide all data to recalculate the SS.

SS calculation is not perfect, since it requires data of previous studies of the same medication or the same intervention that is intended to be studied. It is implicit that the results observed in previous research are going to be the same to those the investigator will observe in his, which is highly unlikely. A degree of uncertainty in the SS calculation is thus introduced, but it is worse performing studies without having an idea, even if it is approximate, of how many cases are needed to consider that an intervention is positive, or else, that it doesn't have any effect.

Indispensable data in protocols' SS calculation

There are four data to be included in controlled clinical trial protocols:

1. *Z alpha*: It is the Z value that corresponds to a given alpha error. The alpha error is falsely rejecting H_0 [null hypothesis]. The term *accuracy* is used to refer to the probability of incurring this error.
2. *Z beta*: It is the Z value that corresponds to a given beta error. The beta error is to falsely accept H_0 . The term *power* is used to refer to the probability of incurring the complement of this error (1 - beta error), which can be seen as the probability of not incurring this error, or else, as the probability of correctly rejecting the H_0 .
3. *VAR*: It is the variance of the outcome variable. The variance is a measure of the variability of a variable and it is calculated differently for numerical ($VAR = \text{standard deviation squared}$) and categorical data ($VAR = p \times q$). The sample will be larger the larger VAR is.
4. *DIF*: It is the minimum difference in the outcome variable that researchers consider clinically relevant. In the study protocols of two groups, control and treatment, the difference of means or proportions is used. The size will be larger the smaller DIF is.

The outcome variable is given different names, including main variable, follow-up variable, dependent variable, but I prefer use *outcome* as an equivalent to *result*, which is consistent with the fact that the outcome variable can be numerical (quantitative) or categorical (qualitative).

Third statement

Equations without both zeds should be considered obsolete and should not be used, since they don't consider beta error Z-value.

As a matter of fact, assessing a lack of effect has become more important than establishing a positive effect. In another words, the rationale of current calculation is opposed to the first SS calculation strategies where the probability for the treatment to be positive was important that and it wasn't even thought about how negative a negative study was. An example of this turn are the strategies to evaluate dose-response studies looking for a SS that allows reaching a 90 or 95 % power and do not take accuracy into account, since the interest is focused on establishing with certainty whether or not there are differences between doses.

Fourth statement

Currently, power is as important as accuracy in SS calculation equations.

Let's see a bit more about the equations' four components:

Z alpha and Z beta depend on the researcher's will within certain limits; for example, it is almost universally accepted that accuracy is between 1 % and 5 % and power between 80 % and 95 %. The range of Z alpha and Z beta values is presented in Table 1 columns 3 and 6, and are used almost by all clinical researchers, and I advise not to leave them out if one wants for a protocol to be approved.

The lowest Z values in the table are those that would yield the lowest SS, but it is recommended not to use lower Z values, especially for power, which has driven various authors to recommend the use of a minimum power of 90 or 95%. See the NOTE in the frame that follows if you want to know why talking about Z-values is talking about probabilities.

Table 1. Range of most commonly used Z-values

Alpha error	Accuracy	Z alpha	Beta error	Power	Z beta
0.1 %	0.1 %	3.29	5 %	95 %	1.65
1 %	1 %	2.58	10 %	90 %	1.28
2 %	2 %	2.33	15 %	85 %	1.04
5 %	5 %	1.96	20 %	80 %	0.84

Note

Any distribution of values, be it Gaussian or not, can be transformed into a Z distribution and when doing it, every Z distribution has Mean = zero and standard deviation = one.

This allows comparing the distributions of different variables because they all are in Std-Dev units instead of the units they were measured with. Consequently, Z values mean the same for any Z distribution. Thus, Z alpha = 1.96 means that 5 % of the area will be left outside at a distance of 1.96 units from the mean zero and, therefore, there will be 5 % of probability for incurring an alpha error in any Z distribution. Therefore, if an accuracy of 5 % is chosen, Z alpha = 1.96 is used, but if another accuracy is chosen, for example 1 %, a Z alpha = 2.58 will have to be used (see Table 1).

Similarly, Z beta = 0.84 leaves 20 % outside the area, i.e., the probability of incurring a beta error is 20 % (80 % power), but if the beta error is wanted to be lower (greater power), Z beta = 1.28 could be used, which corresponds to 90 % power.

In a protocol, including accuracy and the power used in the equation is enough, for example, it was calculated with 5 % accuracy and 90 % power.

VAR and DIF

These are the components the researcher obtains from previous studies where data on treatment response with the same intervention on the same outcome variable are provided. The robustness of these components will be higher the the larger previous information is, and it will be better if there is information on the response in both treated subjects and controls (and not only in treated subjects).

The Z alpha, Z beta and VAR components are always in the numerator of the SS equations while DIF is in the denominator. Therefore, to lower SS, only decreasing VAR or increasing DIF is left to do,

since using Z values below 1.96 and 0.84 is not recommended.

The ideal in a protocol is that it contains a research question that can be answered by measuring the effect of an intervention on a single outcome variable. But if more than one outcome variable is included, then SS should be calculated for each one and use the largest SS of the calculated ones.

Fifth statement

SS cannot be calculated if any of the four calculation components is missing. If there is more than one outcome, the largest SS should be selected.

The only protocols without calculation components are exploratory studies, where there are no previous studies. In such cases, a pilot study with 10 to 20 subjects that serves to calculate the SS is recommended. This strategy is similar to that of adaptive designs, which allow modifying an original protocol, including SS, once the first data of a study are available.

SS in NITs (non-inferiority trials)

Therapeutic advances have made the comparison of new drugs against placebo to be an ethical transgression; therefore, comparing them versus the treatment that is accepted as the best available until that moment and that many refer to as “active control”, to remind us that it will act as if it was the placebo of the control group, is currently recommended.

In NITs, the 95 % CI (confidence interval) of the difference of means or proportions of both groups is calculated. The 95 % CI lower limit of should not exceed the so-called non-inferiority margin (usually symbolized by a lowercase delta, δ). If the 95 % CI is plotted along the X axis, knowing the different results that can occur in NITs becomes easier; for example, Figure 1 shows the 95 % CIs of five studies (A-E) and classifies them according to whether the lower limit of each interval exceeds or not the verticals at zero and $-\delta$. According to this, studies A, B and E are conclusive, whereas the C and D studies need to include more study subjects to clarify whether or not they are inferior to the active control. The investigator chooses the margin (δ) based on his/her judgment and on what is known about the subject in the medical literature. Thus far there are only few recommendations on ways to establish the δ magnitude.³

Figure 1. 95 % CIs of five studies (A-E). New treatment minus active control difference.

TM calculation in NITs

The growing use of NITs deserves noting that the same equations employed in superiority studies can be used, only that the equations' denominator has to be modified. Simply subtract δ absolute value from the difference of means (or proportions) in the equations' denominator:

Non-inferiority denominator = $(\text{DIF means} - \delta)^2$

Superiority denominator = $(\text{DIF})^2$

It can be observed that NITs will always have a lower denominator than superiority studies and, therefore, the SS of a NIT protocol will always be larger than that of a superiority protocol.

Conclusions

The effect size plays an important role in the SS calculation. It is a measure of the strength of a phenomenon and there are very diverse ways to measure it so that a correlation between two variables, a regression coefficient, a difference of means and even the risk of something happening, such as what proportion of subjects will survive a heart attack, are examples of how we measure an effect.

The growing sophistication in the measurement of effect size will lead to new ways to calculate SS.

References

1. Dickersin K, Chalmers I. Recognizing, investigating and dealing with incomplete and biased reporting of clinical research: from Francis Bacon to the WHO. *J R Soc Med.* 2011;104:532-538.
2. Clark T, Berger U, Mansmann U. Sample size determinations in original research protocols for randomised clinical trials submitted to UK research ethics committees: review. *Brit Med J.* 2013;346:f1135.
3. Flight L, Julious SA. Practical guide to sample size calculations: non-inferiority and equivalence trials. *Pharmaceutical Stat.* 2016;15:80-88.