ABC of Methodology

This is a new Section of *Epidemiologia e Psichiatria Sociale*, that will regularly cover methodological aspects related to the design, conduct, reporting and interpretation of clinical and epidemiological studies. We hope that these articles will help develop a more critical attitude towards research findings published in the international literature and, additionally, will help promote the implementation of original research projects with higher standards in terms of design, conduct and reporting.

Corrado Barbui, *Section Editor* and Michele Tansella, *Editor* EPS

An introduction to sample size calculations in clinical trials

SIMONE ACCORDINI

Unit of Epidemiology and Medical Statistics, Department of Medicine and Public Health, University of Verona, Verona, Italy

KEY WORDS: clinical trials, sample size, power analysis, CHAT study.

When planning a clinical trial (Barbui *et al.*, 2007; Cipriani *et al.*, 2007), the investigators must determine how many subjects should be recruited, i.e. the sample size. This is particularly important because studies with too few subjects will not provide reliable answers to the questions addressed (ICH, 1998). Moreover, studies with too large sample sizes may also be unethical, due to the unnecessary involvement of surplus subjects with a consequent increase in costs (Altman, 1980).

Sample size is determined by a statistical calculation that should be performed on a single primary endpoint, which is usually a variable of biological and/or clinical importance, directly related to the primary objective of the trial (ICH, 1998; Chow *et al.*, 2003). The method and the estimates of the quantities used in the calculation should be documented in the protocol and in the study report (ICH, 1996).

The pre-study power analysis is probably the most commonly used method (Chow *et al.*, 2003). According to this approach, sample size is chosen to achieve a desired probability (power) to detect a pre-planned clinically meaningful difference of the primary endpoint between the study groups, at a fixed probability of erroneously rejecting the null hypothesis (significance level). The calculation is carried out by using an appropriate statistical test for the hypotheses of interest, derived under the study design. Besides the primary endpoint, the following items must be specified:

- the null and alternative hypotheses referring to the primary endpoint;
- the clinically meaningful difference to be detected;
- the probability of erroneously rejecting the null hypothesis (significance level) and the probability of rejecting the null hypothesis if the clinically meaningful difference truly exists (power);
- the test statistic.

In clinical trials, a hypothesis is a statement that usually concerns the effectiveness / safety of the treatment under investigation (Chow *et al.*, 2003). In superiority trials, the null hypothesis asserts that there is no difference between the mean response (μ) in the experimental (E) and control (C) groups (H_0: $\mu_E = \mu_C$), whereas the response is assumed to be different under the alternative hypothesis (H_1: $\mu_E \neq \mu_C$). The hypotheses of interest are dissimilar in equivalence trials, which are aimed at demonstrating that the study treatments have no clinically meaningful difference, that is H_0: $\mu_E - \mu_C \leq -d$ or $\mu_E - \mu_C \geq d$ (non-equivalence) *vs* H_1: $-d < \mu_E - \mu_C < d$ (equivalence), d being the largest clinically acceptable difference, and in non-inferiority trials, which are aimed at showing that a given treatment is clinically not inferior as compared to another one, that is H_0: $\mu_E - \mu_C \leq -d$ (inferiority) *vs* H_1: $\mu_E - \mu_C > -d$ (non-inferiority) (Julious,

Address for correspondence: Dr. S. Accordini, Sezione di Epidemiologia & Statistica Medica, Dipartimento di Medicina e Sanità Pubblica, Università degli Studi di Verona, Istituti Biologici II, Strada Le Grazie 8, 37134 Verona (Italy).
Fax: +39-045-505.357
E-mail: simone.accordini@univr.it

https://doi.org/10.1017/S1121189X00002475 Published online by Cambridge University Press

2004). The different hypotheses influence the sample size calculation, as active-controlled trials have a larger sample size than placebo-controlled superiority trials (Hwang & Morikawa, 1999), and non-inferiority trials have a smaller dimension than equivalence trials (Christensen, 2007) and active-controlled superiority trials (Snapinn, 2000).

A clinically meaningful difference of the primary endpoint to be detected in the trial must be provided. The choice of this quantity is particularly important because it strongly affects the sample size calculation. In general, only a few subjects are needed to detect a large difference. In equivalence / non-inferiority trials, both the true difference and the equivalence / non-inferiority limit must be specified, but the setting of the latter is a controversial issue (ICH, 2001; Julious, 2004). When data are normally distributed, the standard deviation of the primary endpoint is also required, and the smaller the variability of the primary variable, the smaller the sample size.

When testing hypotheses, two kinds of errors can occur: the null hypothesis is rejected when it is true (type I error) and the null hypothesis is not rejected when it is false (type II error). In the sample size calculation, the probability of the type I error (significance level α) is controlled at an acceptable level, since this error is usually considered more serious; then the study dimension is chosen to detect the clinical meaningful difference with the smallest probability of the type II error (β) or, equivalently, with the highest power ($1-\beta$) possible, at the fixed α. In general, a conventional choice is 0.05 for the significance level and 0.8-0.9 for power (Chow *et al.*, 2003). When the significance level is fixed, the higher the power, the larger the sample size.

Various test statistics can be used to verify the hypotheses of interest. For example, a z-test or an exact test can be used to test the inequality of two independent proportions. It is very important to choose a test statistic for the sample size calculation whose assumptions will be verified by data, and to use the same test statistic for the analysis of the primary endpoint.

The power analysis performed for the Clozapine Haloperidol Aripiprazole Trial (CHAT) (Barbui *et al.*, 2006) is reported as an example. CHAT is an ongoing randomised, controlled, parallel-group, superiority trial on the effectiveness of clozapine and aripiprazole versus clozapine and haloperidol in the treatment of schizophrenia, with withdrawal from allocated treatment within 3 months as the primary endpoint. On the basis of the data from a recent antipsychotic trial (Lieberman *et al.*, 2005), it has been assumed that the withdrawal proportion within 3 months will be 0.25 (p_C) in the group treated with clozapine plus haloperidol (control group); moreover, it has been hypothesised that the augmentation with aripiprazole (experimental group) will show a clinically significant advantage by producing a withdrawal proportion of 0.10 (p_E). Using the two-sided z-test with pooled variance to verify inequality (H_0: $p_E = p_C$ *vs* H_1: $p_E \neq p_C$) and targeting the significance level at 0.05, a sample size of 194 patients (97 in each group) achieves 0.8 power to detect a difference of 0.15 between the two proportions. Assuming that 10% of the participants could be lost within 3 months or could not provide valid data at month 3, 216 (=194/0.9) patients must be recruited to obtain 194 evaluable patients (Chow *et al.*, 2003). The results of a sensitivity analysis are reported in Figure 1, showing how much the sample size increases if a small difference between proportions must be detected with a high power.

REFERENCES

Altman D.G. (1980). Statistics and ethics in medical research. III How large a sample? *British Medical Journal* 281, 1336-1338.

Barbui C., Cipriani A., Malvini L., Nosè M., Accordini S., Pontarollo F., Veronese A. & Tansella M. (2006). Trasformare la pratica clinica in ricerca. Un invito a partecipare allo studio CHAT. *Rivista di Psichiatria* 41, 326-330.

Barbui C., Veronese A. & Cipriani A. (2007). Explanatory and pragmatic trials. *Epidemiologia e Psichiatria Sociale* 16, 124-125.

Cipriani A., Nosè M. & Barbui C. (2007). What is a risk ratio? *Epidemiologia e Psichiatria Sociale* 16, 20-21.

Chow S.C., Shao J. & Wang H. (2003). *Sample Size Calculations in Clinical Research*. Marcel Dekker: New York.

Christensen E. (2007). Methodology of superiority vs. equivalence trials and non-inferiority trials. *Journal of Hepatology* 46, 947-954.

Hintze J. (2004). *NCSS and PASS*. Kaysville: Number Cruncher Statistical Systems.

Hwang I.K. & Morikawa T. (1999). Design issues in noninferiority / equivalence trials. *Drug Information Journal* 33, 1205-1218.

ICH. E3 (1996). Structure and content of clinical study reports. July 1996. Retrieved July 26, 2007 from http://www.fda.gov/cder/guidance/iche3.pdf.

ICH. E9 (1998). Statistical principles for clinical trials. September 1998. Retrieved July 26, 2007 from http://www.fda.gov/cder/guidance/ICH_E9-fnl.pdf

ICH. E10 (2001). Choice of control group and related issues in clinical trials. May 2001. Retrieved July 26, 2007 from http://www.fda.gov/cder/guidance/4155fnl.pdf

Julious S.A. (2004). Sample sizes for clinical trials with normal data. *Statistics in Medicine* 23, 1921-1986.

Lieberman J.A., Stroup T.S., McEvoy J.P., Swartz M.S., Rosenheck R.A., Perkins D.O., Keefe R.S., Davis S.M., Davis C.E., Lebowitz B.D., Severe J., Hsiao J.K. & Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) Investigators (2005). Effectiveness of antipsychotic drugs in patients with chronic schizophrenia. *New England Journal of Medicine* 353, 1209-1223.

Snapinn S.M. (2000). Noninferiority trials. *Current Controlled Trials in Cardiovascular Medicine* 1, 19-21.

https://doi.org/10.1017/S1121189X00002475 Published online by Cambridge University Press

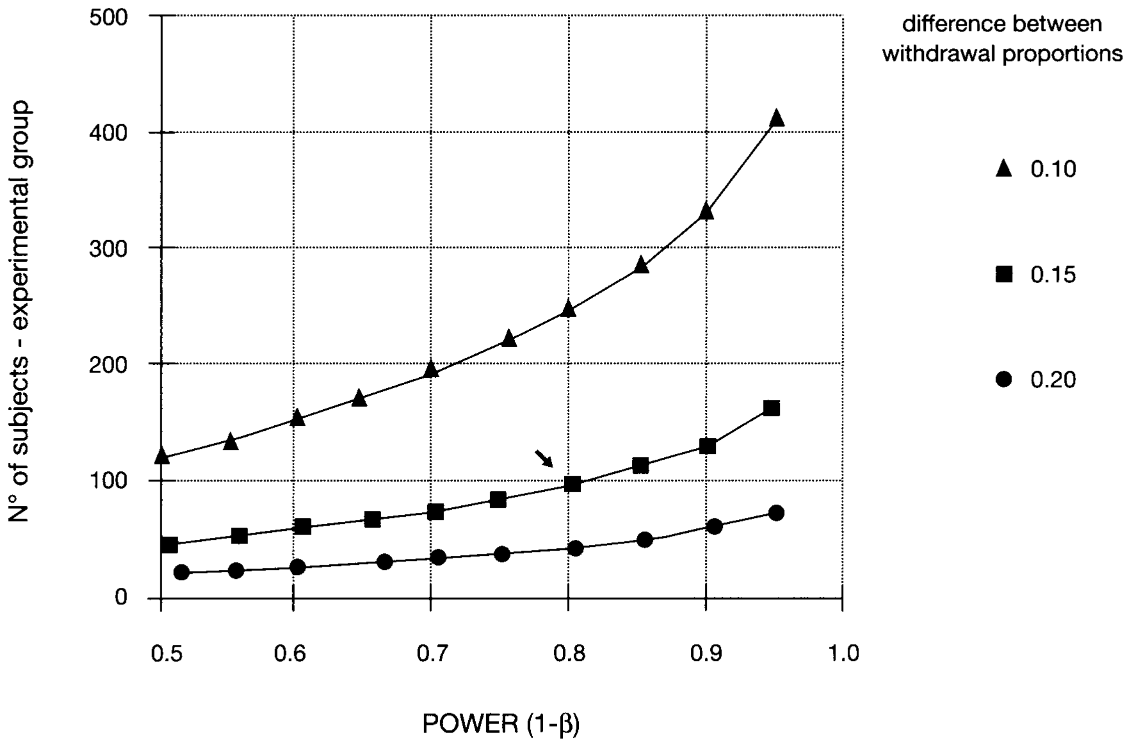

Figure 1. – *Number of subjects to be enrolled in the experimental group, according to different values of power and different assumptions on the difference between the two withdrawal proportions within 3 months. The arrow indicateds the number of subjects reported in the CHAT protocol (without adjustment). The sample size calculations have been performed with PASS software (Hintze, 2004) assuming a withdrawal proportion of 0.25 in the control group and targeting the significance level of the two-sided z-test (with pooled variance) at 0.05.*

https://doi.org/10.1017/S1121189X00002475 Published online by Cambridge University Press