Determination of sample size in clinical trials

The purpose of this problem is to determine the adequate sample size for a multicenter, prospective, randomised, double-blind, dummy controlled parallel study comparing the efficacy and safety of two antibiotics. The discussion will be based on general guidelines for statistics as given in the documents of the European Agency for the Evaluation of Medicinal Products. The usual requirements about accuracy of estimated cure rates and about the power of statistical tests will be presented and the sample size needed to meet these requirements will be determined. As a confirmation, a few simulations will have to be run to establish empirically that the sample sizes determined using mathematical considerations and approximations are correct.

A. Assumptions and theoretical background

Statistical analyses are always based on assumptions about the mechanisms producing the data. It is therefore necessary to clearly state all the assumptions one uses and discuss the rationale for the choices. In this particular study the assumptions are the following.

1. Population: The patients included in the study will be considered a simple random sample from a large hypothetical population of patients suffering from chronic bronchitis. The population is defined by the inclusion criteria in the Clinical Study Protocol. For the purposes of determining the sample size it will be assumed that the population is well defined and there are no systematic biasses in patient selection that could adversly influence the conclusions.

2. Sample: The study will be multicenter, prospective, randomised, double blind, dummy controlled and parallel. It will be assumed that there are no differences between centers so that patients from different centers can be assumed to be sample units from the same population. Also it is assumed that randomisation will be conducted centrally and will not be in any way correlated with other variables like state of health or age. The fact that the study is double-blind and dummy controlled has more to do with elimination of biasses when assessing the patients health and does not have a direct bearing on the sample size considerations.

3. MISSING DATA: An important factor to consider in any clinical study is the question of missing data. There are several reasons for loss of data. Patients may not show up for check-ups, they may not conform to the study protocol, or their attending physician may decide to break off the treatment for medical reasons. The assumption will be made that patients drop out of the study at some fixed rate, INDEPENDENTLY of the treatment or any variables that may be related to the response variable. While this assumption can never be fully justified it is usually taken as the working hypothesis. One has to decide on a case to case basis to what extent this working hypothesis can be trusted. In cases like chronic bronchitis based on the Protocol and similar studies one can safely work with this basic assumption.

4. RANDOMISATION: The usual assumption is that randomisation is done independendently of any other variables. One simply randomly partitions the sample into two equal groups.

5. RESPONSE VARIABLE: The first objective of the study is to determine whether the cure rates for both treatments are comparable. There are thus only two responses: cured and not cured. Patients where it is not possible to decide whether there has been a clinical cure will be classified as missing data so that all assumptions about missing data apply.

6. STATISTICAL TESTS: Statistical tests will be based on cure rates in the treatment and control groups. The conclusions depend on the prescribed smallest effect size Δ. Typically, Δ depends on the expected cure rate and is recommended by various agencies. If the cure rate is in the range between 80% and 90%, a typical choice is $\Delta = 15\%$. When determining the sample size one can consider the effect of the control treatment as known, or one can simultaneously estimate the effects of both treatments. This has a considarable effect on the required sample size. Formulae for the latter case will be considered. The other requirement is the POWER of the statistical test used. One has to be reasonably confident that in case there is a marked difference in efficacy between the two antibiotics, the statistical test will indicate that with high enough probability. The confidence level α will always be taken to be 95% whereas the power β will be prescribed and the sample size chosen in such a way that the required power will be achieved. The

recommended choice for β is 80%. Formulae for sample sizes will be given for general β.

B. Mathematical notation and considerations

Given all the assumptions one can proceed to mathematical considerations. First, the mathematical notation is listed:

1. N will be the total sample size BEFORE RANDOMISATION, i.e. before the group of N parients is randomly divided into two equal groups.

2. After randomisation we will get two groups whose sizes differ by at most 1. Denote the size of the control group by N_C and the size of the treatment group by N_T. At this stage $N_C + N_T = N$.

3. In the control and treatment group we will have some attrition. Assume that tha drop out rate is the same for both groups and denote it by ρ. In the end the two groups for which we have data will be of random sizes M_C and M_T. These two sizes will be binomial random variables with parameters N_C and $(1-\rho)$ for the control group and N_T and $(1-\rho)$ for the treatment group. It is clear that the bigger the rate ρ, the bigger the sample size we need to achive the required power. Sample size formulae for general ρ will be required (typically $\rho = 0.1 - 0.3$).

4. The cure rate in the treatment and control group will be estimated simply by taking the sample rates $\hat{R}_C$ and $\hat{R}_T$. To find the sample size one has to compute (at least approximately) the standard errors for these two estimators. It has to be pointed out that IT IS NOT CORRECT to treat the sample sizes M_C and M_T as fixed. When computing the standard errors the fact that sample sizes are also random has to be taken into account. This will increase slightly the required sample sizes. The unknown "true" cure rates will be denoted by R_C and R_T. For mathematical purposes the rates will not be in percentages but proportions. e. g. $R_C = 0.85$.

5. The quantity of interest in the end will be the difference $\hat{R}_C - \hat{R}_T$. This estimate of the difference in efficacy needs to be compared to the prescribed smallest effect size Δ. One needs to compute the standard error of this difference and then use normal approximation. The accuracy of the normal approximation will be checked by simulations.

6. The confidence level α will be be taken to be fixed (typically $\alpha = 95\%$). Denote the power by β. The power of course depends on the real difference in efficacy which is unknown. The computations will be based on the worst case scenario: if the difference in efficacy is Δ or more, the test has to detect that with probability β or more.

Problems to solve

One of the prerequisites for determining the sample size are formulae for the variances (or standard errors) for the two estimators $\hat{R}_C$ and $\hat{R}_T$. These formulae need to take into account tha random sizes M_C and M_T of the two samples. The following is a brief outline of the mathematical background.

a. If the sample sizes are known, show that the conditional variances of the estimators are given by

$$\mathrm{var}(\hat{R}_C|M_C = m_C) = \frac{R_C(1-R_C)}{m_C}\,, \quad \mathrm{var}(\hat{R}_T|M_T = m_T) = \frac{R_T(1-R_T)}{m_T}\,.$$

b. Argue that conditionally on the sample sizes, the two estimators are unbiased, i. e.

$$E(\hat{R}_C|M_C = m_C) = R_C \qquad \text{and} E(\hat{R}_T|M_T = m_T) = R_T\,.$$

The extreme cases $M_C = 0$ and $M_T = 0$ which never occur in practice are ignored.

c. It is clear that conditionally on $M_C = m_C$ and $M_T = m_T$ the estimators $\hat{R}_C$ and $\hat{R}_T$ are uncorrelated. The formulae for unconditional variances are now obtained by unconditioning. If U is a binomial random variable with parameters n and p, then a good approximation for $E((1/U)1(U > 0))$ is given by

$$E((1/U)1(U > 0)) \approx \frac{1}{np}\,.$$

By unconditioning show that

$$\mathrm{var}(\hat{R}_C) = E\big(\mathrm{var}(\hat{R}_C|M_C)\big) + \mathrm{var}\big(E(\hat{R}_C|M_C)\big)$$

which gives

$$\mathrm{var}(\hat{R}_C) \approx \frac{R_C(1-R_C)}{N_C(1-\rho)}$$

because M_C is binomial with parameters N_C and $(1-\rho)$. Similarly

$$\mathrm{var}(\hat{R}_T) \approx \frac{R_T(1-R_T)}{N_T(1-\rho)}\,.$$

d. To compute the variance of the difference $\hat{R}_C - \hat{R}_T$, the covariance has to be computed. However, the fact that by conditional independence the conditional covariance vanishes and the conditional expectations are constant, it turns out that $\hat{R}_C$ and $\hat{R}_T$ are uncorrelated. Confirm that by using the necessary formulae.

e. Denote $\hat{\Delta} = \hat{R}_C - \hat{R}_T$. Prove that

$$\operatorname{var}(\hat{\Delta}) \approx \frac{R_C(1-R_C)}{N_C(1-\rho)} + \frac{R_T(1-R_T)}{N_T(1-\rho)}\,.$$

It is also reasonable to assume that the difference $\hat{R}_C - \hat{R}_T$ will be approximately normally distributed. Why?

f. Tests will be based on $\hat{\Delta}$. If the purpose is to prove that the treatment under test is no worse then a standard known treatment as in the case of antibiotics one sided tests will be used. Given α the null-hypothesis that the the cure rate in the treatment group is no worse than the cure rate in the control group will be rejected if

$$\hat{\Delta} > z_\alpha \cdot \operatorname{se}(\hat{\Delta})$$

where $\operatorname{se}(\hat{\Delta}) = \sqrt{\operatorname{var}(\hat{\Delta})}$. If Φ is the cumulative distribution function of the standard normal distribution then $\Phi(z_\alpha) = \alpha$. This means that the two treatments will be judged as unequal if the estimated difference $\hat{\Delta}$ is above the threshold $c_\alpha = z_\alpha \cdot \operatorname{se}(\hat{\Delta})$. Comment briefly on this formula.

g. Since R_C and R_T are unknown, and hence the products $R_C(1-R_C)$ and $R_T(1-R_T)$ are unknown we need to choose some plausible values. There are two possible cases:

> Case 1: If approximate cure rates are known, or if plausible expected cure rates are known, use those in the above products. The value of the two products is robust to changes in R_C and R_T so rough guesses are sufficient. In this case the approximate $\operatorname{se}(\hat{\Delta})$ can be computed to determine c_α. When determining c_α one takes $R_C = R_T$. When computing power, one takes $R_T = R_C - \Delta$ as the worst case scenario.

CASE 2: If the cure rates are not known assume the worst case $R_C = R_T = 0.5$. This will ensure that all the requirements about power and confidence levels will be met. It is a conservative estimate that can give bigger sample sizes than necessary. It is desirable to work with rough guesses about R_C and R_T rather than with worst case values.

Comment briefly on cases 1 and 2.

h. Now that the threshold c_α has been determined we turn to power considerations. Let z_β be such that $\Phi(z_\beta) = \beta$ where Φ is again the cumulative distribution function of the standard normal distribution. Given the smallest effect size Δ, the standard error $\mathrm{se}(\hat{\Delta})$ has to be such that

$$P_\Delta(\hat{\Delta} > c_\alpha) = \beta\,.$$

This last equation ensures that the power of the test is at least β if the "true" difference in cure rates exceeds Δ. Why?

i. Finally, the sample sizes can be approximated. The equation

$$P_\Delta(\hat{\Delta} > c_\alpha) = \beta$$

implies that

$$c_\alpha = \Delta - z_\beta \cdot \mathrm{se}(\hat{\Delta})\,.$$

Here one has to be careful about the computation of $\mathrm{se}(\hat{\Delta})$. Again there are two possible cases since R_C and R_T are unknown.

CASE 1: If approximate cure rates are known, or if plausible expected cure rates are known, use those in the above products in such a way that $R_C - R_T = \Delta$.

CASE 2: If the cure rates are not known assume the worst case $R_C = 0.5$ and $R_T = R_C - \Delta$.

We decide for Case 1. Comment why one would do that?

j. This finally leads to a formula for the sample size. The equation in (i) translates into the equation

$$\begin{aligned} & z_\alpha \cdot \sqrt{\frac{R_C(1-R_C)}{N_C(1-\rho)} + \frac{R_C(1-R_C)}{N_T(1-\rho)}} \\ = \quad & \Delta - z_\beta \cdot \sqrt{\frac{R_C(1-R_C)}{N_C(1-\rho)} + \frac{R_T(1-R_T)}{N_T(1-\rho)}}\,. \end{aligned}$$

Note that R_T and R_C are chosen in such a way that $R_C - R_T = \Delta$. Recall that N_C and N_T differ by at most 1 and take $N_C = N_T = M$ as an approximation which has practically no effect on the final sample size. Show that one gets the equation

$$z_\alpha \cdot \sqrt{\frac{2R_C(1-R_C)}{M(1-\rho)}} + z_\beta \sqrt{\frac{R_C(1-R_C)}{M(1-\rho)} + \frac{R_T(1-R_T)}{M(1-\rho)}} = \Delta\,.$$

Denote

$$S_0 = \sqrt{\frac{2R_C(1-R_C)}{(1-\rho)}}$$

and

$$S_1 = \sqrt{\frac{R_C(1-R_C)}{(1-\rho)} + \frac{R_T(1-R_T)}{(1-\rho)}}\,.$$

Solving for M derive, that the sample size is

$$M = \left(\frac{1}{\Delta}\Big(z_\alpha \cdot S_0 + z_\beta \cdot S_1\Big)\right)^2\,.$$

The approximate formula for the TOTAL SAMPLE SIZE is $N = 2M$. Comment on the choices for S_0 and S_1.

k. Take $\Delta = 0.15$, $\alpha = 0.95$, $\beta = 0.80$, and assume approximate cure rates of around 70%. Assume a drop out rate of 15%, i. e. $\rho = 0.15$. Compute the required sample size. Run a simulation to determine whether the power of the test meets the requirements.