AMC Clinical Research Unit

AMC biostatistics manual - Sample size calculation

Nan van Geloven, Marcel Dijkgraaf, Michael Tanck, Hans Reitsma
Clinical Research Unit / KEBB

Academic Medical Center Amsterdam

October 29, 2009

Introduction

This manual provides a practical guide into sample size calculations used in medical research. After reading the manual a medical researcher will know:

- why power analysis is used to plan and evaluate medical research
- what the concepts of ”power” and ”statistical significance” mean
- what information is needed for a sample size calculation
- where to find the information needed
- how to perform a simple sample size calculation
- how to write down a power calculation.

In addition, the manual contains two worked-out examples of sample size calculations in medical research.

Disclaimer
This manual is a practical guide teaching you the basics of sample size calculation; it is not a comprehensive textbook. When in doubt about the correct way to set up and analyze your study, do consult an epidemiologist or biostatistician.

Version 1.0, October 2009
Question and suggestions about this manual are most welcome. Mail us at statistiek@amc.nl.

Contents

1 Why perform a sample size calculation prior to a study?

Why do we want to determine the sample size before a study starts? The importance of a sample size calculation is based on ethical grounds. If the number of subjects tested in a study is too small to pick up a possible effect in the population, study subjects are tested in vain. The study will easily result in a false negative conclusion. On the other hand, testing too many subjects may also lead to undesirable situations. If an intervention turns out to be effective, too many subjects have missed out on this intervention. If the intervention is not effective, too many have been exposed to this ineffective intervention. For these reasons a trial should always consider what number of subjects would be appropriate to answer the study question. Sample size calculations prior to a study can help focus on the number of subjects that is needed and sufficient for a study. Moreover, a sample size calculation helps one to focus on a clinically relevant effect, instead of the erroneous strategy of testing as many subjects as needed to reach statistical significance of an irrelevant effect.

Several regulatory authorities demand a sample size calculation before the start of inclusion of subjects. The CONSORT statement (guideline for reporting clinical trials) states that a researcher should calculate study size on beforehand and should report this calculation in the methods section of the resulting scientific paper. The AMC Medical Ethics Board (MEC) and the Animal Experiments Committee (DEC) also ask for a power calculation in the approval process. The same holds for most study grant applications (e.g., ZonMW). Finally, the logistic planning of a study benefits from a sample size calculation.

2 What is power and statistical significance?

The term 'power' pops up everywhere in medical research, certainly in sample size calculations. Often, the term power is interpreted as a synonym for the number of patients tested in a study. 'Our study did not have enough power to control for possible confounders' is understood as 'you didn't test enough patients to account for several effects'. 'Our study had 80% power to detect an OR of 1.1 at a significance level of 5%' is understood as: 'you have tested enough patients to pick up a possible effect'. Although these interpretations are not (absolutely) wrong, in order to use the concept of power in a sample size calculation, we need to understand its exact meaning. Formally:

> the power of a study testing the null hypothesis H_0 against the alternative hypothesis H_1 is the probability that the test (based on a sample from this population) rejects H_0, given H_0 is false (in the whole population).

So the power is the chance of correctly rejecting a null hypothesis (rejecting a null hypothesis given it should be rejected). Since in most tests H_0 is stated as 'no difference between groups or no effect of intervention', for example H_0 = 'no difference in survival between treated and control group', rejecting H_0 means you have reason to believe there is a difference. In other words, the power reflects the ability to pick up an effect that is present in a population using a test based on a sample from that population (true positive).

The power of a study is closely related to the so called type II error (β), the probability of falsely accepting H_0. The power of a study is $1 - \beta$, so it is the probability of rightfully rejecting H_0 (see Table 1). In the table also the significance level α is stated. Alpha is the probability of falsely rejecting H_0, i.e., falsely picking up an effect (false positive). Note that α only concerns about situations in which no true effect exists in the population.

In a sample size calculation one determines the number of patients needed to test the hypothesis with large enough power and small enough significance level. In this way one protects oneself against false negative and false positive conclusions.

Table 1: Possible conclusions and errors of a study in relation to the truth.

		Whole population	
		effect exists H_1 is true	no effect exists H_0 is true
Study conclusion	effect observed H_1 appears true	true positive power $(1 - \beta)$	false positive type I error (α)
	no effect observed H_0 appears true	false negative type II error (β)	true negative $(1 - \alpha)$

3 What information is needed to calculate a sample size?

To make a sample size calculation based on the power of a study one will need information about each of the following values:

Desired power of the study $1 - \beta$ How much power do you want in the study? Or, stated differently, how certain do you want to be of preventing a type II error?

Desired significance level α How certain do you want to be of preventing a type I error?

Desired test direction One or two sided test?

Clinically relevant (or expected) difference Which difference or which effect are you trying to find?

Expected variance / standard deviation How much variation is expected in subjects belonging to the same study group?

Test to be used in statistical analysis How will the hypothesis test be performed in the analysis phase of the study?

Attrition rate Anticipate on the number of included subjects who will not be available for the study analysis.

4 Where to find the information needed for a sample size calculation?

In this section we advice on how to determine or choose the necessary input values for a sample size calculation.

Desired power of the study 80% is a common power level used in sample size calculations. It means that you accept a chance of 20% (one in five) of failing to detect an effect in your study sample that is indeed present in the population (false negative). If you want to reduce the change to miss out a certain effect, you should increase the power level for instance to 90%. Increasing the power level will increase the sample size.

Desired significance level 5% is a common significance level used in hypothesis testing. This means that you accept a chance of 0.05 to detect an effect in your study that is not present in the whole population (false positive). A reason to lower the significance level might be that multiple tests are done and you do not want to detect an effect just by increasing the odds of finding a false positive. Lowering the significance level will increase sample size.

Desired test direction A two sided test is standard. It means that you test the possibility that treatment A is better than treatment B and the other option (treatment B better than treatment A) simultaneously. A one sided test can only be considered when a clear rationale is provided about why only one direction of the alternative hypothesis is tested (ethical committees and journals are quite strict on this point, some even reject all one sided tests). See e.g. Knottnerus (2001) or Peace (1989) for considerations about using a one sided test for sample size calculation.

Clinically relevant (or expected) difference. Here you have to define the difference that you would like to detect with your study. It can be the effect that has been found in previous studies and that you would like to reproduce. Or in situations where there are no previous studies, you can define a difference that you consider clinically relevant. Information can be found in previous studies found in literature or can be based on expectation from clinical practice. Since a small effect is more difficult to pick up than a large effect, decreasing the difference (or effect) will increase sample size. Note: frequently, available time and resources do not allow the conduct of a clinical trial large enough to reliably detect the smallest clinically relevant effect. In these cases, one may choose a larger difference, with the realization that should the trial result be negative, it will not reliably exclude the possibility of a smaller but clinically-important treatment difference (Lewis, 2000). In general, you have to find a balance between defining a large(r) effect that is easier to pick up (i.e. requiring fewer subjects) and running the risk of obtaining a non-significant result if the difference turns out smaller.

Expected variance / standard deviation. This should be based on pilot data or previous projects in your institute or comparable studies found in literature. When no direct estimate of the standard deviation is available, nQuery (next Section) offers some help tables. If high variation exists between subjects, a difference between groups or an effect of intervention will be harder to pick up, so more spread in the data will increase sample size.

Test to be used in the statistical analysis A power calculation will always be based on one particular statistical analysis. Therefore, the sample size calculation forces you to think about the planned data analysis in a very early phase of the study. Help on the correct choice of analysis can for instance be found on the 'wiki biostatistiek' (http://biostatistiek/mediawiki, available from the AMC network).

Attrition rate Previous studies in the same population will give an estimate of the expected number of included subjects who will not be available for analysis. This may be caused by dropout or withdrawal from the study. Study burden, follow up length and for instance age will influence the attrition rate. The simplest form of attrition, i.e. attrition not related to the intervention or the outcome, can be easily corrected for in the sample size calculation. After calculation of sample size, adjust so that the number needed remains after expected loss of study subjects. For example: if an attrition rate of 10% is expected, divide the number needed by 0.9 (1-attrition rate).

Since one never knows the exact difference, variation or attrition rate of a study on beforehand, a sample size calculation remains a difficult exercise. Repeat the calculation using slightly different input values and check the consequences of these modifications. If absolutely no information is available for the estimation or the necessary input values, one may consider doing a pilot study first.

5 Which software can be used for sample size calculations?

Several computer programs exist for performing sample size calculations. The AMC has a license for the program nQuery Advisor, a user friendly application that supports several study designs and types of data. The software programme can be downloaded directly from the AMC-NAL (migrated computers go to Start - All Programs - Install extra software) or from the CRU website: (http://www.amc-cru.nl/tools.aspx?panel=SOF). For more advanced study designs, KEBB and CRU have access to the program NCSS PASS. On the internet several free power programs exist. Reliability of these free programs is not always ensured. We advise you to use nQuery for your power calculations and contact us when the study design and/or planned analysis is not supported by nQuery (statistiek@amc.nl).

6 Examples

6.1 Comparing 2 means - fever and chronic rhinosinusitus

Suppose we want to investigate the symptoms of patients undergoing surgery for chronic rhinosinusitus (CRS). In particular we are interested whether these patients show signs of fever. We plan to measure pre-operative temperature of CRS patients and of healthy control patients undergoing a cosmetic nose surgery. We want to know how many patients to include in our study to be able to detect a clinically relevant difference between both groups.

From previous studies in which temperature of healthy persons was measured, we estimate that the mean temperature in the control group will be 37^o C, with a standard deviation of 0.4. In our study we are interested in detecting the clinically relevant difference of 0.5^o C, which means we want to be able to detect if the CRS group has mean temperature 37.5^o C or higher. We assume that the CRS and the nose surgery group have comparable spread in temperature (common standard deviation of 0.4).

In nQuery we take the following steps:

1. We go to File → New.

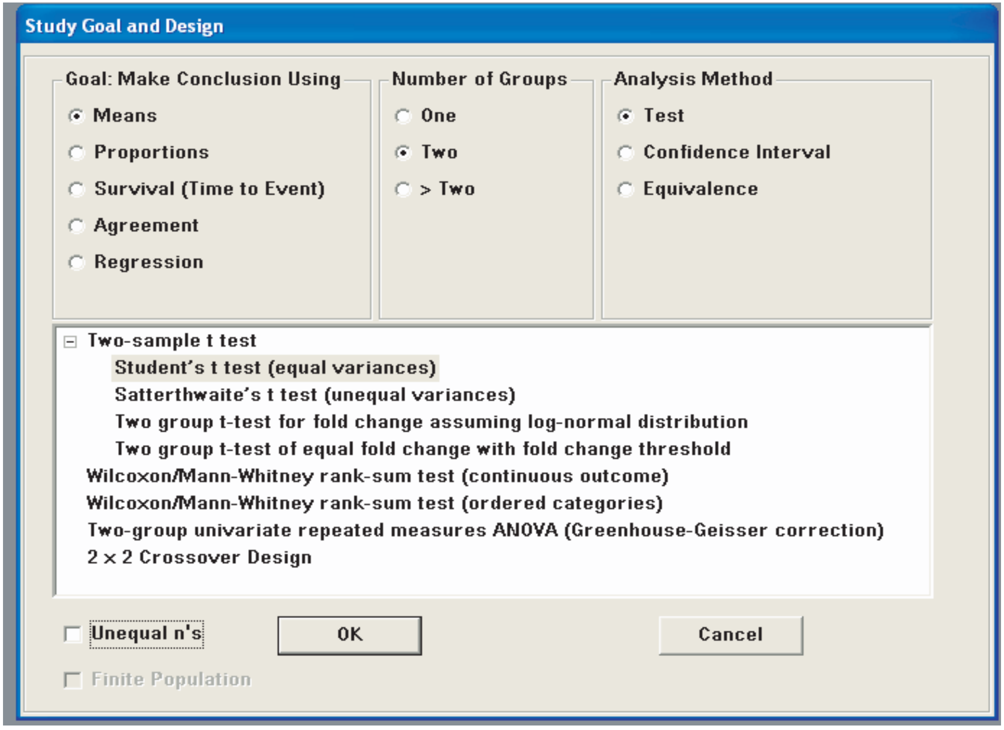
2. We tick that we are comparing two means and that we want to test the difference between these two means using a Student's t-test assuming equal variances (Figure 1).

3. In the next screen we enter our study assumptions: significance level 0.05, 2 sided test, group 1 mean 37, group 2 mean 37.5, common standard deviation 0.4, power 80% and see that nQuery calculates 12 persons per group to be tested (Figure 2).

4. To test the effect of slightly different assumptions on this calculation, we can alter some settings and see that we need 17 persons per group to detect a difference of 0.4^o C or to detect the difference 0.5^o C when standard deviation is higher (Figure 2).

Anticipating on a 10% attrition rate due to invalid measurements, we decide to include 19 ($\frac{17}{0.9} = 18.9$) persons per group, or 38 in total.

Figure 1: Standard settings for comparing two means (nQuery screenshot)



Study Goal and Design

Goal: Make Conclusion Using
- ● Means
- ○ Proportions
- ○ Survival (Time to Event)
- ○ Agreement
- ○ Regression

Number of Groups
- ○ One
- ● Two
- ○ > Two

Analysis Method
- ● Test
- ○ Confidence Interval
- ○ Equivalence

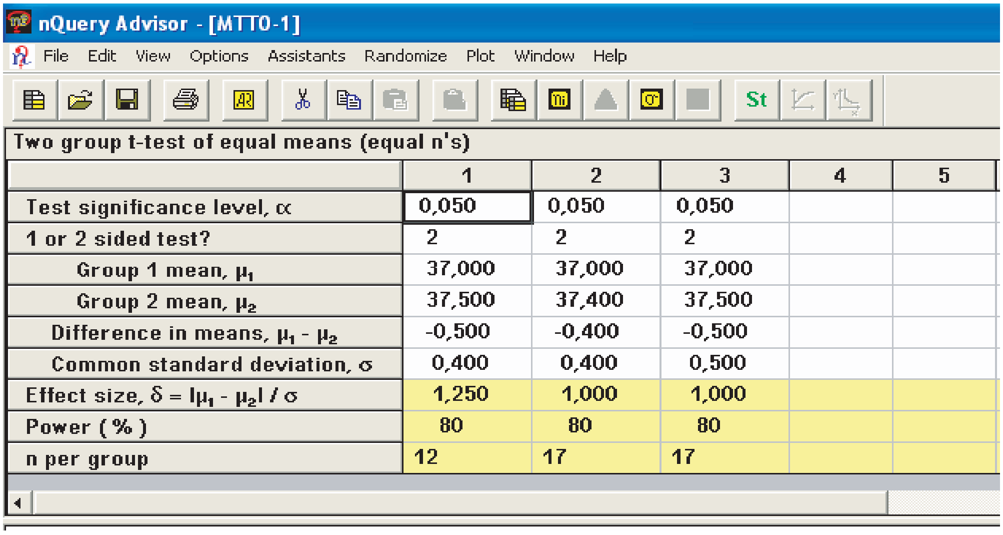
⊟ Two-sample t test
- Student's t test (equal variances)
- Satterthwaite's t test (unequal variances)
- Two group t-test for fold change assuming log-normal distribution
- Two group t-test of equal fold change with fold change threshold

Wilcoxon/Mann-Whitney rank-sum test (continuous outcome)
Wilcoxon/Mann-Whitney rank-sum test (ordered categories)
Two-group univariate repeated measures ANOVA (Greenhouse-Geisser correction)
2 x 2 Crossover Design

□ Unequal n's

□ Finite Population

Figure 2: The assumptions for the CRS study (nQuery screenshot)



nQuery Advisor - [MTT0-1]
File Edit View Options Assistants Randomize Plot Window Help
St

Two group t-test of equal means (equal n's)

	1	2	3	4	5
Test significance level, α	0,050	0,050	0,050		
1 or 2 sided test?	2	2	2		
Group 1 mean, μ_1	37,000	37,000	37,000		
Group 2 mean, μ_2	37,500	37,400	37,500		
Difference in means, $\mu_1 - \mu_2$	-0,500	-0,400	-0,500		
Common standard deviation, σ	0,400	0,400	0,500		
Effect size, $\delta = \lvert\mu_1 - \mu_2\rvert / \sigma$	1,250	1,000	1,000		
Power (%)	80	80	80		
n per group	12	17	17		

6.2 Comparing 2 proportions - methods to help stop smoking

Suppose we want set up a randomized controlled trial to compare two methods of helping smokers to give up smoking. One group is to be given a new kind of nicotine chewing gum and the other group will receive advice from their doctor and a booklet. We want to know how many patients to include in our study. Based on published evidence we expect that in the advice group 15% of smokers will remain non-smokers at 6 months. We would be interested in an improvement to 30% in the group given gum. In nQuery we take the following steps:

1. We go to File → New.
2. We tick that we are comparing two proportions and that we want to test the difference between these two proportions using a Chi-square test (Figure 3).
3. In the next screen we enter our study assumptions: significance level 0.05, 2 sided test, group 1 proportion 0.15, group 2 proportion 0.30, power 80% and see that nQuery calculates 121 persons per group to be tested (Figure 4).
4. To test the effect of using slightly different assumptions and we repeat the calculation and see that we need 250 persons per group to detect a difference between proportions of 0.10 (0.15 versus 0.25) and only 73 per group to detect a difference between proportions of 0.20 (0.15 versus 0.35) (Figure 4). Note: If numbers per group were low, we could also check how using a continuity corrected chi-squared test would alter our calculations. Continuity adjustment to chi-square is usually applied to tables with one or more cells with frequencies less than five. Some authors also apply it to all 2 by 2 tables since the correction gives a better approximation to the binomial distribution. (Garson, G. David, 2009).

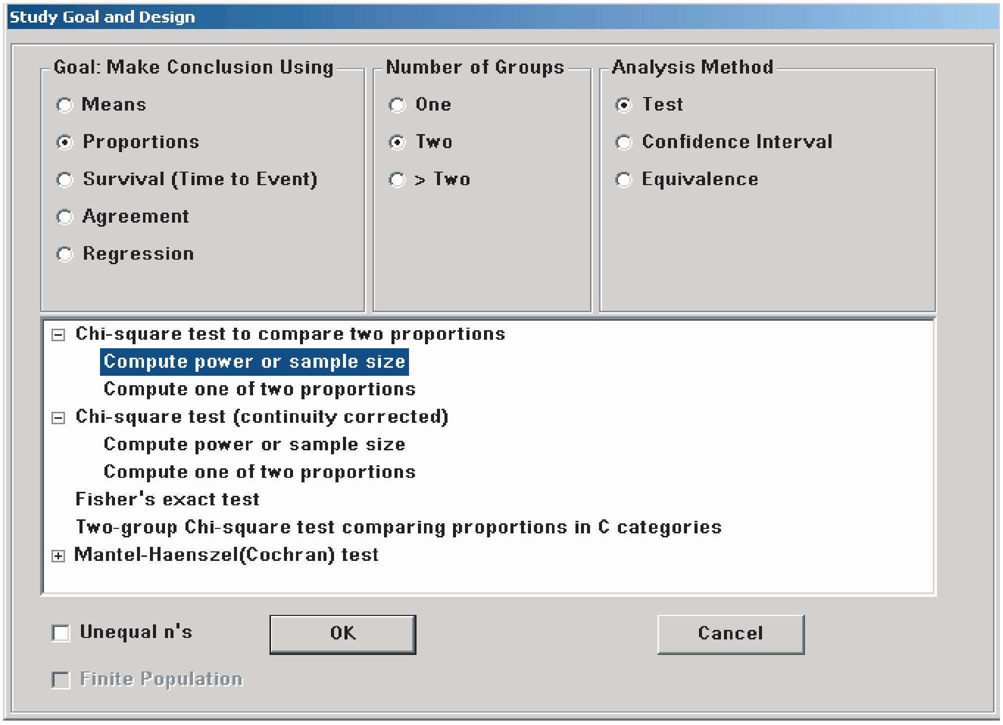
We here accept that our study is not powered on finding a difference smaller than 15% between groups. Anticipating on a 10% attrition rate and keeping our initial assumptions for the difference in proportions, we decide to include 135 (121/0.9 = 134,44) persons per group, or 270 in total.

7 How to write down a sample size calculation?

nQuery can help you writing down your sample size calculation with the 'create statement' function (available under the 'Edit' tab). For the CRS example given in the previous section it will generate the following statement:

> A sample size of 12 in each group will have 80% power to detect a difference in means of -0,500 (the difference between a Group 1 mean, 1, of 37,000 and a Group 2 mean, 2, of 37,500) assuming that the common standard deviation is 0,400 using a two group t-test with a 0,050 two-sided significance level.

Figure 3: Standard settings for comparing two proportions (nQuery screenshot)



Study Goal and Design

Goal: Make Conclusion Using

- ○ Means
- ● Proportions
- ○ Survival (Time to Event)
- ○ Agreement
- ○ Regression

Number of Groups

- ○ One
- ● Two
- ○ > Two

Analysis Method

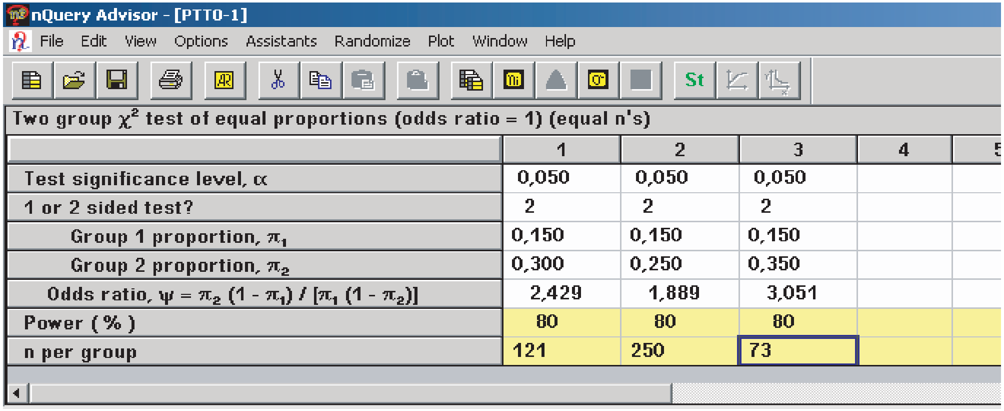
- ● Test
- ○ Confidence Interval
- ○ Equivalence

```
⊟ Chi-square test to compare two proportions
      Compute power or sample size
      Compute one of two proportions
⊟ Chi-square test (continuity corrected)
      Compute power or sample size
      Compute one of two proportions
  Fisher's exact test
  Two-group Chi-square test comparing proportions in C categories
⊞ Mantel-Haenszel(Cochran) test
```

□ Unequal n's

□ Finite Population

Figure 4: The assumptions for the smokers study (nQuery screenshot)



nQuery Advisor - [PTTO-1]
File Edit View Options Assistants Randomize Plot Window Help
St

Two group χ^2 test of equal proportions (odds ratio = 1) (equal n's)

	1	2	3	4	5
Test significance level, α	0,050	0,050	0,050		
1 or 2 sided test?	2	2	2		
Group 1 proportion, π_1	0,150	0,150	0,150		
Group 2 proportion, π_2	0,300	0,250	0,350		
Odds ratio, $\psi = \pi_2 (1 - \pi_1) / [\pi_1 (1 - \pi_2)]$	2,429	1,889	3,051		
Power (%)	80	80	80		
n per group	121	250	73		

When asked to report a sample size calculation this automatically generated statement is a good starting point. You should customize this statement and complete it with references on which you based the assumptions. A paragraph on a sample size calculation is very explicit; a reader has to be able to reproduce your calculations. In the CRS example the paragraph could read like:

> CRS example
>
> A sample size of 17 in each group (or 34 in total) will have 80% power to detect a difference in means of at least 0.5^o C using a two group t-test with a 0.05 two-sided significance level. In this calculation we used the following assumptions: We expect the healthy patients undergoing a cosmetic nose surgery have a mean temperature of 37^o C (ref 1). We assume that both groups show equal variability in temperature and that the common standard deviation is 0.4^o C (ref 2). A mean temperature of 37.5^o C or higher in the CRS patients is considered a relevant sign of fever. We anticipate that only 90% of included patients will have valid measurements. We therefore plan to include 38 patients in total, 19 in the CRS group and 19 in the cosmetic surgery control group.

For the second example on methods to help stop smoking nQuery generates the following statement:

> A two group 2 test with a 0,050 two-sided significance level will have 80% power to detect the difference between a Group 1 proportion, 1, of 0,150 and a Group 2 proportion, 2, of 0,300 (odds ratio of 2,429) when the sample size in each group is 121.

Our paragraph on sample size calculation could now read as:

> Smoking example
>
> A sample size of 121 in each group will have 80% power to detect a difference between proportions of at least 15%, using a two group chi-square test with a 0.05 two-sided significance level. In this calculation we used the following assumptions: We expect that in the control group that gets physician's advice and a booklet, 15% will remain non-smokers after 6 months (ref 1). An improvement to minimal 30% non-smokers after 6 months in the nicotine chewing gum group is considered as relevant. We anticipate that only 90% of included patients will be available for follow up after 6 months. We therefore plan to include 270 patients in total, 135 in the advice group and 135 in the nicotine chewing gum group.

8 Advanced topics

Several situations exist which call for more advanced sample size calculations. Here we point out some of these situations to make you aware of the need for extra effort when the situation occurs.

Equivalence design In an equivalence design you do not want to test for differences , instead you want to show equivalence. In such a design you will need to specify what your interpretation of similar is. Perfect equivalence can never be demonstrated. A limit has to be determined of which small difference between groups will be considered not meaningful and lead to the conclusion of equivalence, this is called the equivalence limit difference. Also the expected difference between groups has to be given. A special type of equivalence designs is a non-inferiority design. In this design one is interested in equivalence in only one test direction. For instance when a new, less invasive diagnostic procedure is compared to the current invasive one, the new procedure does not have to prove better than the current one. If it has at least similar diagnostic strength as the invasive one it would be preferred.

Clustered design When randomization of patients in not done individually but in clusters (e.g., per treating physician or per department), it is expected that the outcome of patients within a cluster are not independent of each other. In the sample size calculation the correlation between patients needs to be accounted for. Also when multiple observations per patient are obtained, the power calculation has to be suited to take along the correlation between measurements in the same patient.

Advanced analyses Planned statistical analyses such as survival analysis, regression analysis, and reliability analysis call for their own specific sample size calculation.

9 Recommendations for further reading

For further information about sample size calculations we recommend the following papers:

S.R. Jones, S. Carley, An introduction to power and sample size estimation, Emergency Medicine Journal, Volume 20, Issue 5, 2003, Pages 453-458

S. Carley, S. Dosman, S.R. Jones, M. Harrison, Simple nomograms to calculate sample size in diagnostic studies, Emergency Medicine Journal, Volume 22, Issue 5, 2005, Pages 180-181

Sally M. Kerry, J Martin Bland, Sample size in cluster randomisation, BMJ, Volume 316, Februari 1998, Page 549.

S.D. Walter, M. Eliasziw, A. Donner, Sample size and optimal designs for reliability studies, Statistics in Medicine, Volume 17, Issue 1, Januari 1998, Pages 101-110

J. Li, J. Fine, On sample size for sensitivity and specificity in prospective diagnostic accuracy studies, Statistics in Medicine, Volume 23, Issue 16, August 2004, Pages 2537-2550

A. Donner, Sample size requirements for the comparison of two or more coefficients of inter-observer agreement, Statistics in Medicine, Volume 17, Issue 10, May 1998, Pages 1157-1168

Karl E. Peace, The alternative hypothesis: One-sided or two-sided?, Journal of Clinical Epidemiology, Volume 42, Issue 5, 1989, Pages 473-476

J. Andre Knottnerus, Lex M. Bouter, The ethics of sample size: Two-sided testing and one-sided thinking, Journal of Clinical Epidemiology, Volume 54, Issue 2, February 2001, Pages 109-110

Roger J. Lewis, Power Analysis and Sample Size Determination: Concepts and Software Tools, Presented at the 2000 Annual Meeting of the Society for Academic Emergency Medicine (SAEM) in San Francisco, California

Garson, G. David (2009) "Chi-Square Significance Tests", from Statnotes: Topics in Multivariate Analysis.
Retrieved 11/06/2009 from http://faculty.chass.ncsu.edu/garson/pa765/statnote.htm.