Special Report Clinics in Orthopedic Surgery 2013;5:235-242 • http://dx.doi.org/10.4055/cios.2013.5.3.235

How to Calculate Sample Size and Why

Jeehyoung Kim, MD, Bong Soo Seo, MD

Department of Orthopedic Surgery, Seoul Sacred Heart General Hospital, Seoul, Korea

Why: Calculating the sample size is essential to reduce the cost of a study and to prove the hypothesis effectively.
How: Referring to pilot studies and previous research studies, we can choose a proper hypothesis and simplify the studies by using a website or Microsoft Excel sheet that contains formulas for calculating sample size in the beginning stage of the study.
More: There are numerous formulas for calculating the sample size for complicated statistics and studies, but most studies can use basic calculating methods for sample size calculation.
Keyword: *Sample size*

In recent years, as the institutional review board has become mandatory, estimation of the sample size has attracted people's attention. Still, many clinicians need to learn why the sample size needs to be calculated and how to calculate it.

It is thought by some researchers that if they conduct a sample size calculation, they need to investigate a high number of samples whereas they only have limited time and money. Some of them even treat it as a kind of rite of passage. Also, they think it is too hard to calculate because they need to use complicated formulas.

Rather, sample size calculation is an indispensable process for obtaining optimal results. Indeed, researchers should know how to calculate sample size because they have limited time and money. Simply, to save time and money, researchers should calculate the sample size.

As researchers usually want to prove that the experimental group is superior to the control group, this article will focus on the superiority trial and we will discuss the non-inferiority trials next time.

WHY

Many researchers want to show that the two groups are truly distinct, but they will fail to find significant differences if the sample size is not big enough. Also, they can waste time and money by continuing an investigation past the time it needs to be continued because they do not know when the testing has been completed since they haven't calculated the sample size before the investigation begins. If the sample size is already large enough to prove that the experimental group is superior, maintaining treatment for the control group could be an ethical problem because the treatment they are receiving is obviously inferior. Thus, it is clear that calculation of sample size is essential ethically and also effectively to get the greatest satisfaction at the lowest cost.

WHEN

Calculation of the sample size is carried out during the planning stage. Thus, calculating the sample size is usually conducted in prospective random control studies. Retrospective studies use statistical power rather than the calculation of sample sizes and we call these 'post hoc power analyses'. We are going to learn about the need and the worth of these 'post hoc power analyses' later.

Also, because researchers expect to uncover findings by referring to previous research studies or pilot studies, the calculation of sample size is done after references are

Received June 14, 2013; Accepted July 15, 2013
Correspondence to: Jeehyoung Kim, MD
Department of Orthopedic Surgery, Seoul Sacred Heart General Hospital, 259 Wangsan-ro, Dongdaemoon-gu, Seoul 130-011, Korea
Tel: +82-2-966-1616, Fax: +82-2-968-2394
E-mail: kjhnav@naver.com

investigated, and before the full-scale research begins.

HOW

The method is pretty simple. First, there is the primary outcome, according to whether the primary outcome is binary variable like pass/fail or a continuous variable like weight/height/score, the methods will be explained one by one in the next section.

When the Primary Outcome Is a Binary Variable

Let's make an assumption that the success rate of the control group and the experimental group is 70% and 85% respectively, as calculated by previous research or pilot study. Visit http://www.sealedenvelope.com/power/binary-superiority/ and click 'calculate' after putting in the success rate which is mentioned above.

As the results show, the sample size required per group is 118 and the total sample size required is 236 (Fig. 1). The statistical significance level, alpha, is typically 5%

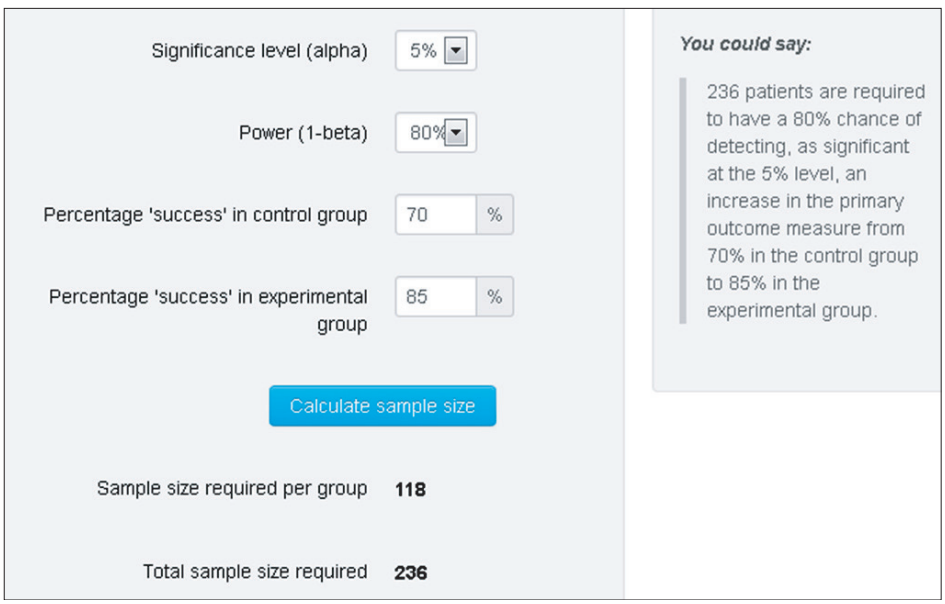

Fig. 1. An example of sample size calculation for a binary outcome superiority trial. Adaped from http://www.sealedenvelope.com/power/binary-superiority/ with permission from Sealed Envelope.

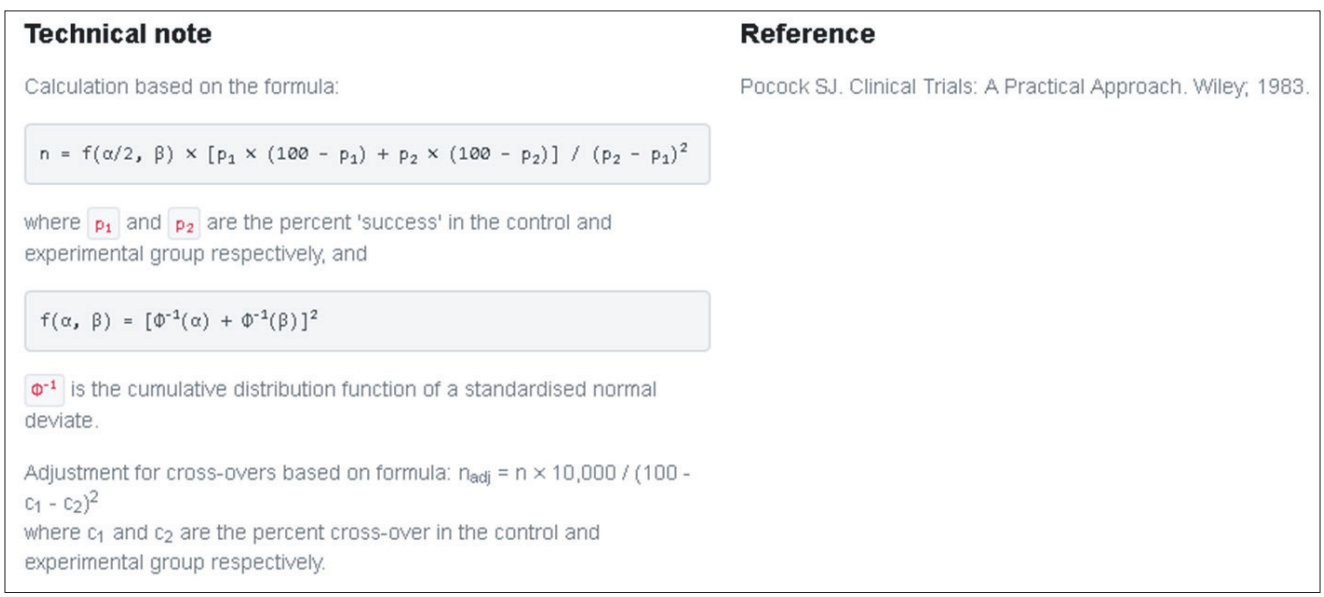

Fig. 2. The formula for a binary outcome superiority trial. Adaped from http://www.sealedenvelope.com/power/binary-superiority/ with permission from Sealed Envelope.

(0.05) and adequate power for a trial is widely accepted as 0.8 (80%). The higher the power (power = 1 – beta) for a trial, the larger the sample size that is required. The right part in Fig. 1, 'You could say~', shows an example of a sentence that can be used in the paper. The meaning of alpha and beta is very important, but it will be left out because it has already been explained precisely in many statistical references.

If you don't know that the sample size required is 236, you will not be able to detect the difference for your inadequate sample size. After you estimate the time which is required for 236 patients, you can change the subject of the study or look for co-researchers or change the dependent variables if the sample size is too large to be collected. Thus, it is necessary to calculate the sample size for estimating the direction of the entire study, as well as the time the study takes, and the budget for the study.

Next, there is a formula for the calculation (Fig. 2). This formula is commonly listed in statistical textbooks and is covered in statistical lectures, therefore, there may be no need for it to be explained. Overall, this is a very simple calculation.

There are many sites for calculating sample sizes. One of them is shown in http://department.obg.cuhk.edu.hk/ and go to the statistical tool box → statistical tests → sample size → compare proportions → independent groups. The same as before, after the deletion of the %, input 0.70 and 0.85. The ratio is 1, usually (Fig. 3).

This result shows that 121 or 134 patients per group are required (Fig. 4). The former is the result of the 'Uncorrected chi-square test', and the latter is the result of 'Fisher's exact-test or with a continuity corrected chi-squared test'. Although the latter is a more accurate way to get the sample size, it does not matter if the former formula is used.

Compared with several other websites, the results of 'www.sealedenvelope.com' are different a little. These differences are thought to be due to the roundings.

Sample size for comparing event rates between two independent Cohorts

How: Place the 2 anticipated proportions/event rates in the appropriate text boxes, and click the Calculate button. The results will show. Note that proportions are entered as a number between 0 and 1, so that 25% is entered as 0.25.

0.70	Estimated proportion in group 1 (controls)	Calculate
0.85	Estimated proportion in group 2 (study)	
1	Ratio controls to experiment subjects	

Fig. 3. Another example of sample size calculation for a binary outcome superiority trial. Adapted from http://department.obg.cuhk.edu.hk/.

Sample size estimates per group for 2 sided test assuming two groups are independent

Assuming outcome data will be analysed prospectively by uncorrected chi-square test

	Type I error = 0.05	Type I error = 0.01	Type I error = 0.001
Power = 80%	121	180	263
Power = 90%	161	229	322
Power = 99%	280	367	483

Fishers exact sample size estimates per group for 2 sided test assuming two groups are independent

Assuming outcome data will be analysed prospectively by Fisher's exact-test or with a continuity corrected chi-squared test

	Type I error = 0.05	Type I error = 0.01	Type I error = 0.001
Power = 80%	134	193	277
Power = 90%	174	242	335
Power = 99%	293	380	496

Fig. 4. The results of Fig. 3. Adapted from http://department.obg.cuhk.edu.hk/.

When the Primary Outcome Is a Continuous Variable

Visit http://www.sealedenvelope.com/power/continuous-superiority/ and input the numbers. The means and standard deviations in the control group and experimental group are required this time.

If the means and standard deviation of the experimental and control group are 76, 83 and 10 respectively, 66 samples (33 samples per each group) are calculated (Fig. 5).

Follow the menu in the web (http://department.obg.cuhk.edu.hk) → statistical tool box → statistical tests → sample size → compare means→independent groups. You can input the difference of two means and the standard deviations of each group, or you can set the ratio, and the result will be same in both cases (Fig. 6).

It is considered that 33 people will be enough to prove the hypothesis.

These two methods for binary variables and continuous variables are common, simple, and easy ways of calculating the size of samples.

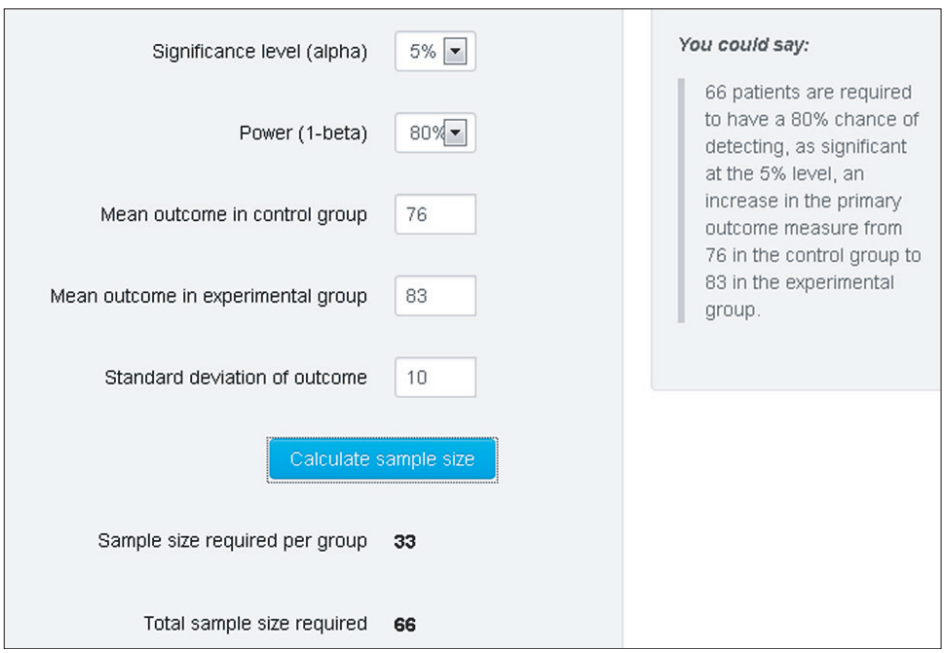

Fig. 5. An example of sample size calculation for a continuous outcome superiority trial. Adaped from http://www.sealedenvelope.com/power/binary-superiority/ with permission from Sealed Envelope.

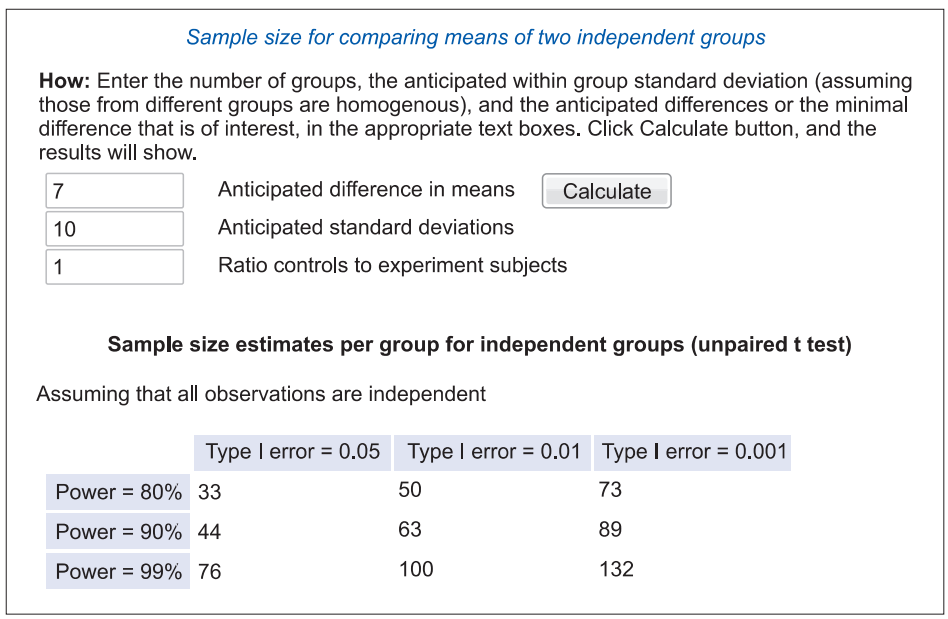

Fig. 6. Another example of sample size calculation for a continuous outcome superiority trial.

Chi-squared test: superiority

Treatment: control ratio	Alpha	Beta	Control rate of success	Treatment rate of success	Sample size	Follow-up loss rate	Compliance	Sample size: Control	Sample size: Treatment
1	0.05	0.2	0.7	0.85	120.471938	0.05	0.9	141	141
			Average rate of success	0.775					

Chi-squared test: equivalence

Treatment: control ratio	Alpha	Beta	Control rate of success	Treatment rate of success	Equivalence limit	Sample size	Follow-up loss rate	Compliance	Control	Treatment
1	0.05	0.2	0.45	0.3	0.1	1951.37942	0.15	0.95	2417	2417
			Average rate of success	0.375						

Chi-squared test: non-inferiority

Treatment: control ratio	Alpha	Beta	Control rate of success	Treatment rate of success	Non-inferiority limit	Sample size	Follow-up loss rate	Compliance	Control	Treatment
1	0.05	0.1	0.45	0.4	0.1	1672.35631	0.15	0.95	2072	2072
			Average rate of success	0.425						

McNemar test: superiority

Treatment: control ratio	Alpha	Beta	Discordant ratio	Treatment rate of success	Sample size	Follow-up loss rate	Compliance	Control	Treatment
1	0.05	0.1	0.5	0.2	131.342788	0.15	0.95	163	163

t-test: superiority

Treatment: control ratio	Alpha	Beta	Standard deviation	Average difference	Sample size	Follow-up loss rate	Compliance	Control	Treatment
1	0.05	0.2	10	5	62.7910379	0.15	0.95	78	78

t-test: equivalence

Treatment: control ratio	Alpha	Beta	Standard deviation	Average difference	Equivalence limit	Sample size	Follow-up loss rate	Compliance	Control	Treatment
1	0.05	0.2	10	3	5	525.371153	0.15	0.95	651	651

t-test: non-inferiority

Treatment: control ratio	Alpha	Beta	Standard deviation	Average difference	Non-inferiority limit	Sample size	Follow-up loss rate	Compliance	Control	Treatment
2	0.05	0.1	30	10	16	321.144276	0	1	322	644

Paired *t*-test: superiority

Treatment: control ratio	Alpha	Beta	Standard deviation of difference	Difference of mean	Sample size	Follow-up loss rate	Compliance	Control	Treatment
1	0.05	0.1	30	10	94.5668076	0.15	0.95	118	118

McNemar test: superiority

Treatment: control ratio	Alpha	Beta	Discordant ratio	Treatment : rate of success	Sample size	Follow-up loss rate	Compliance	Control	Treatment
1	0.05	0.1	0.5	0.625	13.4495015	0.15	0.95	17	17

Control	Treatment	Number
+	+	30
+	−	20
−	+	20
−	−	10

f = 0.5

d = 0.625

Discordant ratio
Treatment : rate of success

		Treatment	
		+	−
Control	+	30	20
	−	20	10

Fig. 7. Author's Excel file for sample size calculation for various tests.

Chi-squared test: superiority

Treatment: control ratio	Alpha	Beta	Control rate of success	Treatment rate of success
1	0.05	0.2	0.7	0.85
			Average rate of success	0.775

Sample size
120.471938

Follow-up loss rate	Compliance
0.05	0.9

Sample size	
Control	Treatment
141	141

Fig. 8. Author's Excel file for sample size calculation; chi-squared test.

Calculating a 'Follow-up Loss'

One further consideration might be the 'follow-up loss'. If the sample size is calculated to be 33 but the follow-up loss is assumed to be about 15%,

(initial sample size) × 0.85 = 33
(initial sample size) = 33/0.85 = 38.8235

In this way, the initial sample size will be 39, considering the 'follow-up loss' and you can mention about these processes at the beginning of the statistical section of the paper. The number of samples is not calculated in the basic SPSS statistical program and you do not have to mention the specific statistical program.

More Complicated Cases

There can be many different situations requiring calculation of sample size. Visit my personal blog page (http://cafe.naver.com/easy2know/6259) and download "sample size calculation. jeehyoung kim". This is a read-only file but all the functions are unlimited. There are brief instructions in the first sheet, and the Korean version is in the second sheet, and the English version in the third sheet (Fig. 7).

In the chi-square test (Fig. 8), the results are the same as Fig. 4. For example, if you input 0.7 in the control group and 0.85 in the experimental group, the incidence density is calculated as 120.472 which are the same as Fig. 4 which shows 121.

The special feature of this formula is that more precise control of 'alpha' and 'beta' is possible, and you can adjust the ratio of the experimental to the control group. If you input a specific value in the 'follow-up loss' and 'compliance' cell, it will be calculated immediately in the next cell.

In the case of surgical trials, the compliance would always be 1, but in the case of medical trials, compliance might be less than 1 due to patient's condition.

This Excel sheet has more formulas for calculation for superiority test, non-inferiority test, equivalence test, the goodness of fit test of chi-square, furthermore, superiority test, non-inferiority test, equivalence test of independent *t*-test, paired *t*-test, McNemar test, and survival analysis (log rank test) (Fig. 8).

FURTHER INFORMATION FOR CALCULATING THE SAMPLE SIZE

Programs

- G*Power: http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/
  G*Power is a representative program for the sample size calculation, but it is hard to use and the instruction manual is also difficult to find. You can refer to "*Sample size calculation* (ISBN 9788994467764)".
- Piface: http://homepage.stat.uiowa.edu/~rlenth/Power/ (free)
- PS: http://biostat.mc.vanderbilt.edu/wiki/Main/PowerSampleSize (free)
- NQuery Advisor: Program for calculating the number of samples (charged)
- R (free), Medcalc (charged), SAS (charged), Expansion of SPSS (charged): Statistical program which can calculate the sample size also.

Websites

- http://department.obg.cuhk.edu.hk
- http://www.quantitativeskills.com/sisa/calculations/samsize.htm
- http://www.statstodo.com/SSizSurvival_Pgm.php
  The links listed above have a function of common statistical solutions as well as sample size calculation.

- http://www.cct.cuhk.edu.hk/stat/Proportions.htm
- http://www.sealedenvelope.com/power/
- http://www.dartmouth.edu/~eugened/power-samplesize.php
- http://cafe.naver.com/easy2know/6259
- http://www.danielsoper.com/statcalc3/calc.aspx?id=5
  The links listed above are only for calculating the sample size.

· http://www.epibiostat.ucsf.edu/biostat/sampsize.html?iframe=tr#regress

This is a generalized portal for calculating sample size.

Useful websites are free and simple to use with detailed instructions. However, the function is limited to basic calculations, and the website can always change. But they are still recommendable.

MORE CONSIDERATIONS

Non-significant Result Means

There are some articles that draw a conclusion that there is no difference between two groups because $p > 0.05$, without calculating sample size. This is clearly a fault because whether a significant difference exists or not, the size of the samples is too small to make a conclusion. Many authors make the same mistakes and researchers warn against this kind of mistake. 'Absence of evidence is not evidence of absence'[1] is a free article which contains practical examples, and I highly recommend it to be read. *Statistics in orthopaedic paper*[2] showed a series of errors in orthopaedic papers; e.g., saying "a non-significant result from a two-sample *t*-test does not imply that the two means are equal, only that there is no evidence to show that they are different."

Indeed, when a survey of 170 orthopaedic papers was conducted in Journal of Bone and Joint Surgery (British), Injury, and Annals of the Royal College of Surgeons of England, 49 papers (28.8%) said that the two groups did not have significant differences but only 3 (6.1%) of the papers calculated the sample size.[3]

If you want to make a conclusion that there is no significant difference, you should perform an equivalence test or non-inferiority test. This will be explained another time.

More Than 3 Groups

The Anova for testing several groups involve a complex calculation process, there is also a reciprocal action consideration. Most high-quality papers focus on the comparison of two groups because a specific goal of proving one hypothesis is more important than simultaneous proof of two or three hypotheses. Therefore, rather than comparing as many groups as possible, we recommend to compare just two groups.

Subgroup Analysis

Usually we do not have to conduct a sample size calculation or a test power calculation for subgroup analysis or secondary outcome like the complication rate, but still we need to interpret the results with the concept of power or sample size. For example, it will be impetuous to say there is no significant difference when the significance of the secondary outcome is larger than 0.05 because subgroup analysis always has a small sample size and it is hard to show a meaningful difference.

Effect Size

The concept of 'effect size', which some statisticians favor, is important but not always used in practice. If you want to use the powerful methods like 'G*Power', there is a need to know 'the effect size' first and then calculation of sample size can proceed.

Cohen's method, in which the 'effect size' is computed as large, medium, or small, is not recommended. It is the last method to use, and only when we do not have any pilot study or previous research as a reference, because it suggests constant sample size even when the character of the study is different. Wikipedia mentions this method in an article.[4]

Unexpected Stop of Study

Although it would be desirable if we can test statistical significance after completing every planned sample, sometimes significant difference can be verified with only a small sample size and unexpected complications occur with significant frequency in the study. We have to make a plan or adjustment to the study when these problems arise because it would be immoral for the investigator to continue the test regardless of that complication.

Intention to Treat and Per Protocol

The process about 'follow-up loss' patients is divided into intention to treat (ITT) and per protocol (PP), and the researcher should mention which process is used in the paper. In the former, the study progresses with the initial allocated number of patients, and in the latter, the study progresses with the number of patients who have completed the whole protocol. When researchers consider complications as the primary outcome, they usually use ITT because ITT is more conservative. When researchers want to find out significant differences of effects, they usually use ITT. So ITT is usually recommended in superiority trials and PP and ITT in non-inferiority of effect. Actually, if the result values of these two methods are different, it means many follow-up losses exist. In that case, we should investigate the reason precisely.

For Smaller Sample Size

If authors want to prove the hypothesis with a small

sample, there are some tips such as: 1) Use continuous variables rather than nominal variables; 2) Reduce standard deviation by precise and exact estimation of the continuous variable; 3) Use a statistical matching method if proper (like paired *t*-test); and 4) Set common and distinct variables as primary outcomes.

Blood pressure (continuous variable) is better than hypertension (nominal variable), and if you measure the blood pressure exactly, you can reduce the sample size by decreasing the standard deviation. The paired *t*-test and McNemar test need a smaller sample size than the independent *t*-test and chi-squared test. If the difference between side-effects is more prominent than that of effects, you can prove the thesis with a small sample by focusing on side-effects. All of these can be controlled by closely analyzing pilot studies and previous research studies, so I want to emphasize the importance of the pilot study again.

CONCLUSION

To predict the results of a study and to complete the investigation successfully, calculation of sample size is an essential process and has many benefits. You can conduct this calculation simply by analyzing previous research, using the results of pilot studies, and using a few websites.

REFERENCES

1. Altman DG, Bland JM. Absence of evidence is not evidence of absence. BMJ. 1995;311(7003):485.
2. Petrie A. Statistics in orthopaedic papers. J Bone Joint Surg Br. 2006;88(9):1121-36.
3. Sexton SA, Ferguson N, Pearce C, Ricketts DM. The misuse of 'no significant difference' in British orthopaedic literature. Ann R Coll Surg Engl. 2008;90(1):58-61.
4. Wikipedia. Effect size [Internet]. Wikipedia; 2013 [cited 2013 May 22]. Available from: http://en.wikipedia.org/w/index.php?title=Effect_size&oldid=554920483.