# Power and precision in research

Angie Wade
Professor of Medical Statistics
Centre for Applied Statistics Courses, Population Policy and Practice Programme, UCL Great Ormond Street Institute of Child Health, 30 Guilford Street, London WC1N 1EH.
awade@ucl.ac.uk

*Are adolescents with constipation more likely to suffer psychological maladjustment?*[1]

*What percentage of Chiari I-type headaches show improvement after foramen magnum decompression (FMD)?*[2]

*Does BCG vaccination reduce early childhood hospitalisation in Denmark?*[3]

*Is diagnosis of coeliac disease associated with differences in adolescent anthropometry?*[4]

*Does visual feedback affect the rate of chest compressions?*[5]

These are all questions asked in recent issues of this journal. In each case the authors collated information from a sample of individuals to yield an answer to their question. Differing study types were used ranging from observational audits and surveys through to randomised parallel and crossover trials. The study designs, participants, settings, sample sizes and key statistics are summarised in table 1.

Table 1: Description of the five studies

| Study | Design | Participants | Setting | Sample size | Key statistics |
|---|---|---|---|---|---|
| 1 | Cross-sectional survey | 13-18 year olds | 5 schools | 1697<br>(114 constipated) | 33.3% of constipated and 14.5% of non-constipated had maladjustment:<br>OR 2.94 (1.95, 4.45) |
| 2 | Retrospective review of audit database/ Before-after study | Chiari I-type headache cases having FMD | Tertiary hospital | 39 | 80% showed improvement post FMD |
| 3 | Randomised controlled trial | Newborn babies | 3 Danish hospitals | 4262 children<br>(2129 BCG,<br>2133 controls) | 1047 hospitalisations BCG group vs 1003 controls. HR 1.05 (0.93, 1.18) |
| 4 | General health examination database | 17 yr old Israeli jews | Eligibility assessmen t for military service | 2,001,353 adolescents;<br>(10,566 with Coeliac Disease (CD)) | Boys with CD had lower BMI (average 21.2 vs 21.7)<br>CD girls were shorter (161.5 vs 162.1 cm on average) |
| 5 | Randomised crossover trial | Hospital staff | Tertiary hospital | 50 pairs of measurements – with and without visual feedback | Rate of chest compressions was lower and less variable in those receiving feedback |

Despite these differences, the same basic principle is followed for each. A sample of the relevant group of individuals is identified and from observing what happens to this sample, inferences are made about the wider population. The inferences may be beneficial to similar individuals and those involved in their care. For example, clinicians trying to determine whether to perform FMD[2] or parents considering the pros and cons of BCG vaccination. [3]

How well a question is answered by the study depends on how large a sample was studied in conjunction with other factors such as the variability of the measurements and/or event rates. Both researchers and patients intuitively understand that findings based on a larger sample are likely to be more accurate, and will have more confidence in results based on a randomised trial of 1,000 individuals than if only 10 patients had been recruited. What is less intuitive is that if a treatment is not shown to be effective in a small sample, it may still have benefits. Similarly, we are generally less inclined to also apply such intuitive logic to observational descriptive studies.

This article will explain the rationale behind consideration of sample size for all types of research study, wherever data is collated to address a proposed research question.

## *Making statements based on the available sample*

To explore these concepts further, consider the Chiari I-type headache study[2]. Here 32/39 of individuals showed improvement post FMD, which gives a sample estimate of 82.1% improvement rate. A 95% confidence interval for the rate is (67, 91%), which means that we are 95% confident that at least 67% of the population (in this case, Chiari I-type headache sufferers) will improve. If we wanted a more accurate estimate, then a larger sample size would give this. For example, suppose 1000 had been assessed and of these 821 had improved, then this would still be a rate of 82.1%, but the 95% confidence interval would now be (80, 84%), and we would therefore be 95% confident that at least 80% would improve.

The data that we have is compatible with population estimates anywhere within the 95% confidence interval at the 5% significance level. Any hypothesised population estimate that we test against outside that interval will yield a p-value $< 0.05$ indicating a statistically significant difference and non-compatibility. There is a conceptual distinction between significance tests, which determine compatibility with a pre-specified hypothesised value, and confidence intervals, which have no prior hypothesis and seek only to identify a range of compatibility. For some study types one approach may be favoured over the other. Significance tests are generally considered when presenting the results of a randomised trial, whereas an observational study may omit this approach entirely depending on the precise research question.For example, suppose that prior to the study the authors ascertained that improvement in at least 75% of patients would warrant routine use of FMD. When testing against a hypothesised value of 75%, the sample of 39 would have yielded a p-value $>0.05$, a non-significant difference, indicating compatibility with 75% (as also shown by the confidence interval which contains this value). For the sample of 1000 (where the 95% confidence interval does not contain 75%), a significant result would be obtained ($p<0.05$), indicating non-compatibility.

This example illustrates the correspondences in interpretation between p-values and confidence intervals. Larger samples lead to narrower confidence intervals as they enable more population scenarios to be excluded. Note that if interest lies on the difference between two groups, for example the heights of 17 year olds with and without coeliac disease, then the hypothesised population value will be zero (ie. no difference in average height between the two groups). The confidence interval will be for the difference in means (or percentages) between the two groups. If this interval excludes zero then the difference is statistically significant.

## *What do power calculations have to do with this?*

Studies don't always give a definitive, or even very useful, answer to the research question posed. For example, if the aim of the headache study had been to determine whether the improvement rate was at least 75%, then the sample of 39 would not have been sufficient. Even though performing the significance test gives a non-significant p-value >0.05, this does not necessarily mean that the improvement rate is lower than 75%. The 95% confidence interval (67, 91%) shows that the data are in fact compatible with an improvement rate as high as 91% (as well as values below 75%). We still don't know whether FMD yields the necessary improvement rate to warrant usage as the answer given from the available sample is too imprecise. If a larger sample had been studied then the results would have been more conclusive.

Before starting a study a decision needs to be made as to whether it is feasible to collect a sufficiently sized sample to address the research question within a plausible time frame. Power calculations can tell us how many subjects are needed to obtain a significant p-value if there is a difference of a given size from the hypothesised value (75% in example above). Alternatively, power calculations can give the number needed to estimate a specified size of estimate (again 75% in above example) with given precision.

## *Finding the right number*

There are formulae, that are known as power calculations, which can be used to identify how many need to be sampled to make the estimates sufficiently precise and/or statistical significance likely.

These formulae can readily be identified in appropriate statistics textbooks and articles,[6-13] or via the many online calculators available, identified by a simple google search. Hence in theory this is a relatively easy process – locate a formula or online calculator, plug in a few values and get a number of individuals/items/things that you should collect information on to answer your specified research question. However, there are decisions to make as to which formula and/or online calculator to use, what values to plug into this and how to properly interpret the number that magically comes out. The aim of this short paper is to guide the reader through some calculations so that they are better equipped to address these issues. Formulae are given and

these relate to those found on our online calculators webpage.[14] (http://tinyurl.com/samplesizeCASC ).

## Using a formula/online calculator

There are two types of sample size formulae. The first of these is aligned with significance testing and providing p-values, the second is akin to confidence intervals. Although within the results you will probably present both of these, you need only do one type of sample size calculation. It depends on whether the emphasis in your study is on detecting a difference, or on estimating parameters with sufficient precision.

In this paper, both types of formulae are presented for both binary and numeric outcomes. A binary outcome is one which takes one of two values. For example, the headache study had a binary outcome - headaches either improved or not and the study was interested to identify the percentage who improved[3]. By contrast, a numeric outcome is usually summarised by the mean of the values. For example, the average of the changes in rates of chest compressions for a clinician using two feedback methods.[5]

Often the population parameter being estimated is the difference between two distinct groups of patients (with and without disease; treated vs control). In this case it is the difference in percentages or means between groups for binary and numeric outcomes respectively that is of interest. The difference may be between distinct groups of individuals, for example the percentage difference between those with and without constipation[1], or the difference in mean BMI between boys with Coeliac disease or not[4]. Alternatively, measurements may be paired, within individual or by taking matched pairs of individuals, and these differences summarised, for example the rate of chest compression when the same clinician uses visual feedback or not.[5] Formulae are shown for these scenarios.

Only formulae for binary outcomes and means are given in this paper. Whilst these are the most commonly used and widely applicable examples of power calculations, it should be noted that formulae exist for many other scenarios, such as non-normal data, hazard ratios, odds ratios and/or hierarchical data. The principles of these formulae are similar although they relate to more complex situations, and are hence beyond the scope of this paper.

## Terms to understand

If the outcome is numeric, then an estimate will be needed of the *variability* of the measurements. This is expressed as the *standard deviation* (*sd*). Larger values of the *sd* indicate greater variation.

With a numeric outcome, we can express the difference we would like to detect as a *standardised difference (sdiff),* which is the difference divided by the *sd*. For example, if the *sd* of the change in rates of chest compression using two methods is 13, a difference of 8 will be equivalent to a *sdiff* of 0.6 (8/13).

*Power* is the ability of a study to detect a difference if it exists, and is usually set at 80, 90 or 95%. A power of 80% means that if the difference exists there is an 80% chance that this study will identify it. Hence there is a one in 5 (20% = 100-80%) chance of not identifying the

difference, so although a power of 80% is often used (and will require smaller numbers), it is generally better to have a higher power.

*Significance level* is usually set at 5%. This is the chance of falsely declaring a difference when the population value is actually as hypothesised.

*Precision* is a measure of how closely we would to estimate the true value. The width of the confidence interval is synonymous with precision. When using the confidence interval based formulae, the precision needs to be expressed in the same terms as the outcome (percentage or mean range) and is defined, in the formulae presented here (and associated weblink given above) as half the width of the resultant confidence interval. For example, to estimate the difference in average height of coeliac and non-coeliac 17 year olds to within ± 0.1, the *precision* is entered as 0.1 and we expect to obtain a confidence interval of width 0.2

## *Some power calculations*

Some commonly used formula are presented here with examples. As explained above, the appropriate formula depends on whether it is a difference that is to be detected or a precision attained, and whether the outcome is numeric or categoric. The formulae are organised accordingly.

### Detecting a difference

1) Numeric outcomes

To detect a specified difference *sdiff* with 90% power between two groups at the 5% significance level requires $\frac{\mathbf{21}}{\boldsymbol{sdiff^2}}$ in each group.

Reduce this number by a quarter for 80% power and add a quarter for 95% power.

For example, the *sd* of heights of 17 year olds is known to be about 10cm and a study aims to detect a difference in average height of coeliac and non-coeliac 17 year olds of 0.5 cm or more. This is a *sdiff* of 0.5/10 = 0.05.

The sample size required to do this is $\frac{21}{0.05^2} = 8400$ per group ie. a total of 16,800 17 year olds, half of whom have coeliac disease.

To detect the same difference with 80% power, would require 6300 per group (8400 x 0.75). For 95% power, the numbers per group need to be increased to 10,500 (8400 x 1.25).

Paired data

If there are pairs of observations and it is the mean difference between pairs that is to be compared to an average of zero (implying no difference), then half the sample size given above is the number of paired observations that need to be made. Note that the *sdiff* must be based on the *sd* of the within pair differences, which will be different to the *sd* within each group as the pairing removes variation due to differences between groups (such as differences in age, sex, diet, exercise level, disease status).

For example, with the rate of chest compressions crossover trial, patients are measured using different methods but it is the difference that is of interest ie. each individual contributes one difference to the dataset despite having 2 measures made. To detect a difference between methods of 0.4 *sdiff* with 90% power at the 5% significance level will require $\frac{10.5}{0.4^2} = 66$ clinicians to make paired measurements. To detect the difference with 80% or 95% power requires 50 and 83 clinicians respectively.

2) Binary outcomes

The percentages in each group with the outcome are to be compared. If these are $\%_1$ and $\%_2$, then this difference can be detected with 90% power at the 5% significance level if the following number are assessed in each group:

$$\frac{\mathbf{10.5}\{\%_\mathbf{1}(\mathbf{100}-\%_\mathbf{1})+\%_\mathbf{2}(\mathbf{100}-\%_\mathbf{2})\}}{(\%_\mathbf{1}-\%_\mathbf{2})^\mathbf{2}}$$

As before, this number reduces by a quarter for 80% and increases by a quarter for 95% power.

For example:

i) The randomised trial of BCG vaccination aimed to detect a fall in the hospitalisation rate of 20% to 16% or lower. For 90% power, 5% significance, this requires:

$\frac{10.5\{20(100-20)+16(100-16)\}}{(20-16)^2} = \frac{10.5(1600+1344)}{4^2}$ = 1932 per group,

a total of 3864 randomised to 2 equal sized groups (BCG vaccinated or not).

ii) Assuming 15% of normal children and 25% of children with constipation have psychological maladjustment, this difference can be identified with 90% power at the 5% significance level with two groups of 331 children ($\frac{10.5\{25(100-25)+15(100-15)\}}{(25-15)^2} = \frac{10.5(1875+1275)}{100} = 331$). For 80% power the sample could be reduced to 249 per group.

Notice that the above two examples relate to quite different study forms, a RCT and an observational study, yet the sample size formula required is the same. The formula required depends on the outcome (in this case difference in percentages in two groups) and the power and significance required.

## Being sufficiently precise

Divide the quantity stated by the *precision* required squared (ie. *precision*$^2$) to give the sample size required to estimate with 95% confidence.

1) Numeric outcomes

**8*sd*$^2$** per group

For example, to estimate the difference in average height between coeliac and non-coeliac 17 year olds to within ± 1cm ( *precision* = 1), assuming a *sd* of 10 cm, will require $\frac{8\ x\ 10^2}{1^2} = 800$ per group, total 1600.

For a more precise estimate within ± 0.5 cm, requires a larger sample of $\frac{8\ x\ 10^2}{0.5^2} = 3200$ per group, total 6400.

Paired data

For paired measurements, the number of pairs required is half of the sample size per group as given above (ie. **4*sd*$^2$** paired measurements).

For example, to estimate the average difference between rates of compression using 2 methods to within ± 4, where the *sd* of the within pair differences is 13, requires $\frac{4\ x\ 13^2}{4^2} = 43$ clinicians making paired assessments.

2) Binary outcomes

To estimate a single percentage ( $\%_1$ ): **4 $\%_1$ (100 - $\%_1$)**

For example, if 80% of patients with Chiari I-type headaches improve then this can be estimated to within ± 8% with 95% confidence using a sample of $\frac{4x80(100-80)}{8^2} = \frac{320\ x\ 20}{64} = 100$ patients.

To estimate a difference in percentages ($\%_1$, $\%_2$) : **4 {$\%_1$ (100 - $\%_1$)+ $\%_2$ (100 - $\%_2$)}**

For example, to estimate a fall in percentages hospitalised when given BCG vaccination from 20% to 16% with a precision of 2.5% (ie. confidence interval width 5%) would require a sample of $\frac{4\ x\ (20.80+16.84)}{2.5^2} = \frac{11776}{6.25} = 1885$ per group, a total of 3770 children.

## *Unequal groups*

Sometimes when two groups are to be compared it is not anticipated that they will be of equal size. For example, in the study to compare psychological maladjustment rates between adolescents with and without constipation, this was a survey across schools and the numbers in the two groups would not be expected to be equal.

An imbalance can be adjusted for by increasing the overall sample size.[7] If the above formulae estimate that *n* per group is required assuming equal sized groups, this is a total sample of 2*n*. To account for an imbalance between groups of *k*:1, the total sample size will need to be increased to $\frac{\boldsymbol{n(1+k)^2}}{\boldsymbol{2k}}$

For example, in the constipation study it was anticipated that about 10% of the children would have constipation. This is an imbalance of 9:1 ie. for every 9 that are healthy, there will be 1 who is constipated and hence k=9.

The previous formula showed that 249 per group were required to detect a difference of 10% in psychological maladjustment (15% and 25% per group) with 80% power and 5% significance. This is a total sample of 598. To adjust for the smaller numbers of constipated children, the sample needs to be increased to $\frac{249\ x\ (1+9)^2}{2\ x\ 9} = \frac{24900}{18} = 1384$, which will consist of 138 with constipation (10%) and 1246 without.

## *Some points to note*

Each of the formulae requires estimation of some quantities, for example the sd of the measures or the percentages in each group. This may appear nonsensical. For example, if we knew the percentages of BCG vaccinated and non-vaccinated who were hospitalised (to put into the formula), we would not be doing a study to estimate the percentage difference between the groups!

Sample size estimation is not a precise art. It can give guidance and ball-park figures, but if all information for exact calculation were available we would not need to do the study.

This does not mean that sample size calculation is not worth doing. It is important to ensure that a study is likely to have a reasonable chance of yielding useful information. It is unlikely that we would have no idea of the likely range of estimates required and these can be used to inform calculation.

The size of difference to be detected, sdiff or the difference in percentages, should be informed by an understanding of the minimal clinically important difference. It is important that all estimates are clinically plausible and suitably justified, with estimation never determined with reference to the available or preferred sample size.

In the above calculations, the formulae give the minimum number required and in all cases, where this is not a whole number, has been rounded up. The number given is the minimum

number for statistical analyses, consideration needs to be given when designing the study to the proportion of eligible participants that will refuse to take part and the likelihood of missing data and/or loss to follow up. These factors will impact on the feasibility of attaining the necessary sample size in the timeframe available.

There are some relatively minor discrepancies with the sample size estimates given in the example papers and the numbers given here, but this does not indicate errors. All calculations agree to a reasonable extent, with differences attributable to differing approximations in the formulae used.

There are often other factors that should be considered. For example, if we compare heights between those with and without coeliac disease, we may wish to adjust for differences in age, sex and ethnicity of children as well as other potential confounders associated with anthropometry. It is worth noting that in this case, taking into account these covariables will only serve to reduce overall variation (*sd*), so estimation not taking them into account will be conservative. The formulae given are based on a single primary outcome. If there are multiple comparisons, then the sample sizes will need to be larger. As noted earlier, only the most basic calculations are given in this paper, further formulae exist for more complex scenarios.

No amount of increasing the sample size can account for biases in the data.

## *"But I'm not doing a clinical trial so this is irrelevant"*

It's a common misconception that power calculations are only required for clinical trials of the randomised controlled type variety. This is untrue.

Wherever sampled data is used to address a research question, the sample size needs to be adequate to give a useable answer to that question.

In this paper, a variety of examples are given to illustrate applicability. The headache study was a single sample descriptive study, the differences in constipated and non-constipated an observational comparison.

## *Conclusions*

This article should provide a large enough sample of information to enable the reader to understand that sample size calculation should always be given consideration before commencing a study with 95% confidence.

It should be borne in mind that having an adequate sample size is only one important facet of research design. Samples selected should be representative and measures of outcome both reliable and valid. No amount of increasing the sample can correct for such deficits in design and doing so will merely result in obtaining a more precise estimate of the wrong answer, or the right answer to the wrong question.

When preparing a manuscript for publication this journal refers authors EQUATOR guidelines[15] which aim to ensure that whatever type of study is being undertaken all important statistical considerations are made. The majority of the guidelines make reference to sample size determination, including those for observational (STROBE[16]) and diagnostic testing (STARD[17]) studies. This paper explains the basis of sample size calculation with simple examples given as manual calculations, replicable via online calculators if preferred. Sample size estimation need not be overly complex nor a black-box affair. The aim of this paper is to enhance adherence to guidelines and incorporation of this important element of research in practice.

## References


1. Ransinghe N, Devanarayane NM, Benninga MA, *et al.* Psychological maladjustment and quality of life in adolescents with constipation. *Arch Dis Child* 2017;102;268-273.
2. Raza-Knight S, Makard K, Prabhakar P, *et al.* Headache outcomes in children undergoing foramen magnum decompression for Chiari I malformation. *Arch Dis Child* 2017;102;238=243.
3. Stensballe LG, Sorup S, Aaby P, *et al.* BCG vaccination at birth and early childhood hospitalisation: a randomised clinical multicentre trial. *Arch Dis Child* 2017;102;224-231.
4. Assa A, Frenkel-Nir Y, Leibovici-Weissman Y, *et al.* Anthropometric measures and prevalence trends in adolescents with coeliac disease: a population based study. *Arch Dis Child* 2017;102;139-144.
5. Gregson RK, Cole TJ, Skellett S, *et al.* Randomised crossover trial of rate feedback and force during chest compressions for paediatric cardiopulmonary resuscitation. *Arch Dis Child* 2016;0;1-7.
6. Chow S-C, Shao J, Wang H. Sample size calculations in clinical research 2nd edition. Chapman and Hall CRC Biostatistics series. 2008
7. Altman DG. Practical Statistics for Medical Research. Chapman and Hall/CRC. 1999
8. Campbell MJ, Machin D, Walters SJ. Medical Statistics: A textbook for the health sciences. 4th edition. Chapter 14: Sample size issues. John Wiley and Sons. 2007
9. Wade A. Study Size. *Sex Transm Inf* 2001; 77; 332-334.
10. Bland M. An Introduction to Medical Statistics 3rd edition. Chapter 18: Determination of sample size. Oxford Medical Publications. 2000.
11. Senn S. Statistical Issues in Drug Development. Chapter 13: Determining the sample size. John Wiley and Sons Ltd. 1997.
12. Julious SA. Sample Sizes for clinical trials. Chapman and Hall CRC Press. 2010
13. Kirkwood BR, Sterne JAC. Essential Medical Statistics 2nd edition. Chapter 35: Calculation of required sample size. Blackwell Science Ltd. 2008
14. Wade AM, Koutoumanou E. Centre for Applied Statistics Courses (CASC). Sample size calculation excel sheets [cited 2017, Apr 27] Available from: http://tinyurl.com/samplesizeCASC .

15. Enhancing the QUAlity and Transparency Of health Research. EQUATOR network. [cited 2017, Apr 27] Available from: http://www.equator-network.org/
16. Strengthening the reporting of observational studies in epidemiology. STROBE statement. [cited 2017, Apr 27] Available from: https://strobe-statement.org/index.php?id=strobe-home
17. STARD 2015: AN updated list of essential items for reporting diagnostic accuracy studies. [cited 2017, Apr 27] Available from: http://www.equator-network.org/reporting-guidelines/stard/