3.2 Power calculations

When designing a clinical study a key decision is how many patients (i.e., study subjects) to include. This is required for planning, budgeting and ethical approval. Patients responses to the treatment will be monitored and quantified using at at least one key clinical parameter. If they respond positively such that (for example) that parameter goes up in a desirable way, then the amount it rises relative to baseline levels can indicate the magnitude of the treatment effect. However, this is not the end of the story.

Patients can get better or worse in the absence of any treatments. Similarly, the chosen clinical parameter will show some degree of random variation, which will be evident when the placebo group data is examined. This random variation could be due to measurement errors, measurement method variation or other sources of random noise and systematic biases. When we perform statistical testing we seek to show that the treatment effect is larger than can be accounted for by that random variation in the data.

When the treatment works, can we see that it works?
Suppose we *know* that we really is effective and it will improve our parameter of choice. The true change in the parameter is usually termed the *signal*. How can we demonstrate that the signal exists if it is swamped by a much larger amount of random variation? We can't, as any statistical tests would return a false negative result; the signal, the treatment effect is therefore invisible to us.

Of all the things that affect random variation there is one factor not mentioned above, but which is often tractable and within our control. That is the number of measurements we make or, put another way, the number of individual patients enrolled in the study. The number of patients is of fundamental importance. If we have two groups of otherwise similar people, one group with a thousand patients and one with ten, and make measurements in both groups then the observed data variation (calculated as either variance or standard deviation) will be far less in the large group than the small. The message is clear: large patient numbers result in lower random noise than small patient numbers, Consequently, to see a very small signal (a small treatment effect) then a large number of patients must be included in the study; a big treatment effect can be seen even with low patient number. This is described as the large study having higher *power* than a small study. But how many patients would give sufficient power? To find this this we need power calculations.

A hypothetical clinical trial
Consider first an imaginary study we have already completed called the DOSH trial, which compared a treatment group to a control (placebo) group. The DOSH trial investigated whether a therapy

makes people richer; 'richer' was measured this by determining the amount of money they had in the bank, i.e., their bank balance or **BB,** on a certain date. The trial aimed to determine whether the therapy affected the magnitude of **BB**, and if so, by how much**.**

The importance of being significant

As noted above, demonstrating that a therapy effect on **BB** depends on showing the differences between the treatment and control group wealth level are unlikely to be due to random variation. For this the DOSH trial used a t-test analysis of the mean **BB** levels of the two groups. DOSH did indeed observe an increase with wealth of treated people and showed that the standard deviation of the data was small enough that the **BB** data from the treated group had a p-value of 4.5%. This was regarded as statistically significant since it was below the 5% significance threshold. The study author inferred that the therapy really did increase **BB** levels, so was confirmed as a lucragenic treatment.

Job done.

Power and what it means

Now suppose that we will do another imaginary study of the same treatment on a different population, called the MoOLA trial. How many individuals do we need to recruit to MoOLA so that we will again see a significant difference? This question relates to the power of the new study. Thus, if there are too few people recruited then it will be difficult to see a difference at the 5% significance level (i.e., p<5%). In this case the study is *underpowered* so you are wasting everyone's time since an effect will not be detected even if it is there. The ethics committee will not take kindly to it either.

If there is a massive abundance of subjects investigated the study may be *overpowered*, which is far less of a worry (and rarely happens), but is also wasting resources and the ethics committee will again be displeased.

How to estimate study power

To work out the number of patients needed to power the MoOLA study to see a statistically significant outcome we need some prior information about that key parameter. Since we have the data from the earlier DOSH study we can use that. We first decide:

1) How big is the effect size we want to detect in MoOLA – do we want to be able to detect an increase of $10 or would we be happy to be able to detect increases of $100?
2) How much **BB** is likely to vary by; for this we use the results of the DOSH study.
3) We choose a significance threshold (here 5%)
4) We choose a tolerable false positive rate, usually 80%.

This is enough to calculate **n**, the number of subjects needed for the study. To do this with the information above if we define *Diff* as the mean difference between treated and control, *SD* as the observed standard deviation (the measure of noise) and *SES* the standardised effect size. SES is used to indicate how big the treatment effect is that we want to detect – SES is calculated as effect size divided by standard deviation.

The power calculation can be done by software or, given the information given above we can simply calculate:

$$SES = \frac{Diff}{SD} \quad \text{and} \quad \mathbf{n} = \frac{16}{(SES)^2}$$

Note that number on the right ('16') is calculated from statistical formulae using the (commonly used) false positive rate of 80% and the p-value threshold selected as5%. If different values for false positive rate and p-value are used this will need to be recalculated with a different formula.

The need for prior data

Notice that for MoOLA we had prior data (DOSH) to help us with the power estimate, since we can use the SD for DOSH as an estimate for MoOLA. However, it is possible that no study of our therapy has been done before and we have to use some other means to estimate SD and desired effect size. However, the prior data only has to provide an estimate, so may be derived from a different but related treatment or may be a guesstimate; some thought needs to go into what might be appropriate. Small (i.e., underpowered) pilot studies might be alright for the purposes of this estimation. Also note that the power or **n** estimate only relates to measuring parameter **BB**; if other parameters are also of interest then the power and **n** may need to be calculated accordingly.

Significance thresholds and the pain they cause

It is important to note that the 5% significance threshold used above is reasonable but is both arbitrary and quite low. The assumption that random chance effects do not exist if their p-value is below 5% is commonly made, but is not very robust. The sad truth is that many studies give conclusions that are weak because of this, but in reality there may just not be enough study subjects to improve the power as much as we want.

Other types of studies

Note that the above description relates to a small clinical trial, but the same or a related statistical approach may be used to estimate power in other types of studies, such as observational and time series data or animal studies; the power calculation method will differ accordingly. In addition, we have assumed here that the t-test for significance is used for normally distributed continuous data. There are many other types of tests (and distributions) which can require power calculations, however, these are beyond the scope of the current article.

Author: Julian Quinn ***Version*** 2.2 (Dec 2020)
Thanks to Dr Richard Piper and Dr Fenglian Xu for reviewing and critiquing this article.

Royal North Shore Hospital, DIVISION of SURGERY and ANAESTHESIA