Methodology (13) Power analysis: the magic of statistics - Or: the difference between significance and relevance

Harald Walach (https://harald-walach.info/methodology-for-beginners/) - May 2014

Normally, the average citizen and average scientist is satisfied when he or she we hear that a research result was "statistically significant". We then commonly mean: the hypothesis with which one started the research is proven, the fact that one is investigating is proven. And conversely, if no significant result is found, we believe that the phenomenon in question has not been found, i.e. it does not exist. This is why, for example, the average doctor, journalist and citizen believes that bioresonance is proven to be ineffective and homeopathy is a placebo, and half of America takes lipid-lowering drugs for the primary prevention of heart attacks, because they believe this is a scientifically proven fact. In this blog I want to take a closer look at some of these opinions and show why they have arisen and ask how justified they are. It will turn out: it has to do with what I call the magic of statistics. That is the question of how powerful a statistical test is. It's related to the question of how big an effect we're studying. And it depends on how big the sample is that we need to really make the effect statistically visible, or to get a significant result. In other words, *if there* is a systematic effect, no matter how large it is, then it can be proven with a study, provided we have enough resources. The question that every reader of a scientific study should ask is not: Is a study significant? But: Is the effect shown, whether significant or not, clinically and systematically important? If it is significant, then we can assume scientific confirmation. If it is not significant, we have to ask ourselves: was the size of the study suitable to find the effect? or vice versa: how large would a study have to be to be able to statistically confirm an effect of the magnitude found with a reasonably satisfactory degree of certainty? This is the essence of the power analysis we are now dealing with.

So in every scientific investigation we are dealing with the interplay of a total of four variables that depend on each other like the parts of a filigree mobile. If we change one, all the others change too. These are:

1. The error of the first kind or the alpha error.
2. The error of the second kind or the beta error.
3. The size of the effect, or the effect size.
4. The size of the study or the number of people studied (in the case of clinical or diagnostic studies) or the number of observations.

1. *The alpha error*:

In terms of content, it is the kind of mistake we make when we claim there is an effect somewhere when it is not there. So, for example, if we claim that homeopathy or bioresonance therapy is effective and in fact it is not, we are making an error of the first kind or an alpha error. The convention of research is usually that we are only willing to accept such an error 5 times out of 100. That is why many studies set the significance level at alpha = 0.05. So if a study finds $p < 0.05$ and the researcher concludes the result is significant, this means, in words: If, on the basis of the study, we claim that the intervention studied is effective, or that the effect found is present, or that the diagnostic studied is selective, then we make a mistake 5 times out of 100 cases.

That is why we also say: the probability of error is 5%. We can also say: with a 5% probability, such a result can come about by chance.

That is why we, as scientists, will insist on raising the alpha level in particularly important cases, let's say to 1%. Then we only make such an error in one out of 100 cases. Accordingly, an effect that becomes visible at a significance level of $p < 0.0001$ is afflicted with a very small alpha error, or the probability that we make a mistake when we claim such an effect exists is not very large. Or: such an effect only occurs very rarely by chance.
Science, as an expression of society's collective effort, is interested in guarding against making false claims. It would do so if it were to claim an effect where none exists. This is why all scientists, journal editors and public opinion attach great importance to controlling the alpha error. After all, they would then be advocating a possibly dangerous, costly or otherwise elaborate procedure, even though there is no factual reason for it. That is why almost all people know about the meaning of statistical "significance", or in other words, the meaning of the alpha error is reasonably clear to everyone.
This is also related to the logic of statistical testing, which goes back to Fisher, among others [1]: Here, a hypothesis is formulated in the form: Treatment x is not different from control treatment y. Homeopathy is as effective as placebo, for example. This is the so-called "null hypothesis", i.e. the assumption that there is no difference or no effect. Now one can find out via the statistical procedure of testing such a hypothesis whether this is probably right or wrong. To do this, one sets up an alternative hypothesis, e.g.

"Homeopathy is better than placebo" and now examines the empirically found difference to see whether it is compatible with the original null hypothesis, i.e. that there is no difference. The procedure for this is a statistical test. I will discuss what exactly lies behind this in another blog. Let's assume for the sake of simplicity that our investigation showed that the probability that the alternative hypothesis is true, i.e. that there is a difference between homeopathy and placebo, is $p > 0.05$, i.e. that we make a mistake in more than 5 out of 100 cases when we claim such a difference. Then the statistical test rejects our alternative hypothesis and conservatively tells us we had better keep the null hypothesis and assume that there is no difference between say bioresonance therapy and placebo. Then they say: there is no scientific evidence of a difference between bioresonance and placebo. Does that mean there is no difference? Not necessarily, because we still have to take into account the three other variables. Because it could be that we are committing an error of the second kind, a beta error.

2. *The beta error*:

This consists of overlooking an effect that is present, that is, falsely claiming that an effect is not present. So if we were to say that bioresonance therapy is a placebo therapy, when in fact this claim is false, we would be making an error of the second kind, or a beta error. While the alpha error reflects the conservative side of science, i.e. the concern not to make false claims, the beta error reflects the lack of sensitivity of a study, i.e. the lack of care in looking, the overlooking of effects because the instruments used do not resolve finely enough. E.g. the fact that one did not know for a long time that infections were caused by bacteria was, technically speaking, a beta error made because of this, because they didn't have high resolution microscopes, just as the ignorance of viruses as possible causes of disease before the invention of the electron microscope was a beta error because you couldn't see them or detect them.

So the beta error is related to overlooking effects. Now the point is this, as can be easily understood intuitively: We can set the alpha error arbitrarily. The beta error is intimately related to the size of the effect under investigation. We can say as much as we like: we do not want to miss any cause of disease. If we don't have an electron microscope or an immunological assay, we can't see the tiny viruses and say there is no cause of disease. We then make a beta error when in reality viruses are disease triggers but we cannot detect them due to lack of instruments.

Similar to the beta error in a clinical trial: we like to overlook effects that are smaller than expected, or smaller than can be made visible with the available study size. Does that make them irrelevant? The example of bacteria that can be seen with a light microscope and viruses that can only be seen with an electron microscope or an immunological assay shows that this is obviously wrong. Because viruses can be just as significant as bacteria and bacteria can cause just as dangerous diseases as an injury with a knife that you can see.

The beta error is thus intimately related to the size of the effect to be expected or investigated. We can at best say: we want to miss an effect, if it exists, at best with a probability of also 5% or 10%, i.e. in 5 or 10 cases out of a hundred we make a mistake if we claim that the effect does not exist at all, i.e. that bioresonance is ineffective. But can we formalize this in the same way? Yes and no. No, insofar as it cannot be done independently of other variables. Yes, insofar as we have to include the expected effect size in our calculation. That is why we have to account for the effect size or effect strength (ES) to be examined.

3. *The effect size*

Effect sizes are numerical measures that indicate how much two groups differ, e.g. a treated and an untreated group, or a group treated with the right drug and a group treated with a placebo (in principle, one can also indicate a correlation between variables, i.e. a correlation coefficient, as an effect size; since this is rather unusual in clinical research, I will not go into it now). One then speaks of an effect size that indicates the difference *between* two groups, referred to in English as "between-group effect size (ES)". You can also quantify the size of the effect you see when you study just one group at two different measurement times, between which something has happened, e.g. a natural change when it comes to maturation and growth, or an intervention. That's then an effect size *within* a group that varies from before to after, or from pre to post.
and is referred to in English as "within-group ES". Such an ES within one and the same group can be determined in a numerically similar way, but it clearly has a different systematic meaning, because it does not quantify a difference between groups, but within one and the same group. Therefore, it is no longer of importance for the further considerations here, and whenever I speak of "effect size" in the following, I will firstly more often use the abbreviation ES for "effect size" and secondly mean the ES *between* two groups.

Two families of effect sizes:

What is also important, not in principle but for concrete understanding, is the fact that there are two fundamentally different effect sizes. This has to do with the fact that events in this world can be divided very roughly into two categories: those whose occurrence or non-occurrence can be determined, and those that can be measured more precisely. We also call events that either occur or do not occur dichotomous. This is because they can be quantified with a yes/no decision: Either it rains, or it doesn't. Either someone is alive, or he is dead. Either someone is cured or still sick, has achieved a certain clinically defined outcome or not. Such characteristics are captured with variables that can be coded with 1 or 0 instead of "yes" or "no". That is why they are called dichotomous variables, a special case of a so-called nominal scaling, i.e. a representation in which numerical values do not have a numerical meaning, but a specially defined meaning in terms of content. In this case, "1" for "event established", whatever the event is, and "0" for "event not established".
Such outcome representations are relatively common in medicine, for example in cancer research when survival is at stake, or when one wants a robust outcome measure to provide clarity to the treating physician about what percentage of a group of patients is are "cured" in the sense of a certain definition. Effect sizes of this family are basically expressed by ratios. For example, by the ratio of events in a group in relation to the non-occurrence of such events in this group and this ratio in turn in relation to the same proportions in the comparison group, i.e. numerically RR = a/n1 : a'/n2. RR is the so-called rate ratio. It indicates how the ratio of improved (a) to the total number in the treatment group is to the same ratio in the control group (a'; n2).

If RR = 2, this means: twice as many people got well in the treatment group as in the control group; analogously, an HR (hazard ratio) of 0.5 means that only half of all adverse events, e.g. death or heart attack, occurred in the treated group [2]. Slightly differently defined measures of this family are the odds ratio (OR), or the logarithmisized odds ratio (log-OR). I will discuss effect size determination separately in another blog.

A fundamentally different family of effect sizes comes from variables that are continuously are measured, so-called "interval-scaled" variables. The best-known interval-scaled variables are temperature or time. Each distance between the variables measures a similarly large interval and assigns numbers to the ratios. In medicine or psychology, people also try to make such measurements as often as possible, because they are intuitively more precise than dichotomous characteristics. For example, the statement: "In Berlin it is 5°, in Singapore 28° C" is more precise than the statement "In Berlin it is cold, in Singapore it is warm". Typical examples of interval-scaled measures are blood pressure, which is measured in mm of mercury column (Hg), or heartbeats per minute, or the time a signal or reaction takes in milliseconds, or pain intensity in mm of a visual analogue scale (VAS). Psychometrics, the art of translating complex constructs such as quality of life, depression, anxiety, etc. into numerical values, can also make such content measurable. This is typically done by psychometric scales that are well constructed and empirically validated.
Effect sizes from variables in this category, usually referred to as "d" - abbreviated for "difference" -, can be obtained by subtracting the mean values of two groups from each other. Because all possible measurement methods measure on very different scales - height is measured in cm or mm, time in seconds, years or milliseconds, pain often in mm of a VAS, quality of life, intelligence or depression in scales without actual naming - one has to come up with a trick to make them comparable. This is the statistical
Trick of "standardization". One achieves this by dividing the difference by the standard deviation of the corresponding measures.

Because mean values and standard deviations of certain measurands are always in a certain relationship to each other, comparability is gained in this way.

The corresponding basic formula is thus **d = m1-m2/sd** [3], where m1 is

the mean value of one group,
m2 is the mean value of the other group and
sd is the standard deviation.

So we can now understand in principle how the size of an effect can be captured in a study. For our context, it is now important to understand: If we have defined a certain alpha error, usually 5%, then our ability to detect an effect depends on how large the effect is. Because the bigger it is, the easier it is for me to see it. To see that a knife causes a wound, I only have to open my eyes. To see how a bacterium infects a cell, I need a light microscope. To see how viruses convert the genetic material of a cell, I need immunological methods. So if I have a large effect, I can make the presence of the effect so visible with relatively few people that a mistake is quite unlikely. But if the effect is not quite so large, I obviously need to use a larger sample. This brings us to the fourth important variable:

4. *The sample size*

So our ability to make a therapeutic effect visible in a clinical or diagnostic study or to separate two groups depends on how large the effect is. This effect size defines how large a sample I need to make the effect visible. The sample size is something like the fineness or coarseness of our instrument. A small sample is comparable to seeing with the naked eye. A relatively large sample is as if we need an aid, glasses or a microscope for example, to see something. A very large sample is like using a high-resolution tool like an electron microscope, an immunological sample or a radio telescope to make a very small effect visible. Does this then mean that one could also save the effort? This question can only be answered clinically and not statistically. The answer to the question depends on how important the effect is. If we are looking at a relatively easy-to-treat disease or one that gets better on its own, like the common cold or flu, then we will only be interested in a really big effect. If we have an extremely dangerous or untreatable disease in front of us, like an aggressive cancer or severe pain, then even a relatively small effect will matter.

The statistical power [4]

The power of a test or statistical power is the probability of detecting an effect with a study of defined size, if it is present. It is defined as 1-beta, where beta is the beta error or second kind error. So if I am willing to accept that I will miss an effect 10 times out of 100 if it is present, then I am potentially making a beta error of 0.10. Accordingly, the power would be 1-beta = 0.9 or 90%. However, as we have seen above, unlike the alpha error, I cannot simply define the beta error and thus the power. Rather, both are intimately related to how large the effect is, that I want to show.

This is because it determines how many people I need to include in a study. If I have a large effect, then for the same alpha error the number of study participants needed is relatively small, even if I want to run little risk of missing the effect and thus require a large statistical power of my test. If the effect I expect to see is relatively small, then for the same alpha error I need to recruit a much larger number of people into my study in order to also keep the beta error small and thus see the effect with a reasonably reasonable statistical power. Figuratively speaking, the smaller the effect, the finer the instruments to see it and the greater the effort that has to be made to keep the visibility the same.
It would be a mistake to say - namely a beta mistake - you can't get sick from bacteria just because you can't see them with the naked eye. It was also a mistake to say that there were no planetary systems outside our solar system (a claim that was once heard often about 20 years ago) just because we didn't have telescopes powerful enough. And it would also be a mistake to say that viruses do not exist just because they cannot be seen with light microscopes.
These are all examples of beta errors. They all demonstrate: the smaller the effect, the greater the effort we have to make. How big the effect is is determined by reality. Because the effect size is an empirical quantity, just as the magnitude of a pathogen is an empirical quantity. Either we determine it. That is usually done with pilot studies. Or we start from a reasonable clinical assumption and consider how large the effect must be in a concrete case to make it worth the effort.

The power analysis determines how much effort we have to put into making an effect of a defined size visible. It tells us how many patients we need if we want to make an effect of a defined size visible in a study and thereby avoid an alpha error of a certain size - i.e. usually only falsely claim a difference in 5 out of a hundred cases - and at the same time only want to miss an effect with a certain probability, i.e. not fall below a beta error of a certain size, usually 10 to a maximum of 20%. In other words: with a defined probability of 80-90% to actually see the effect, if it is there.

This results in the very general rule:

Any effect, no matter how large, if it is real and systematic, can be made visible with a large enough sample.

It goes without saying that "sample size" is synonymous with "money and effort". This is because the cost of a study is directly proportional to the number of patients needed.

The size of the sample needed therefore depends on this,

- how big the effect is
- how sure I want to be that I am not making false claims about an existing difference, i.e. of the size of the alpha error, i.e. the chosen significance level.
- how sure I want to be that I am not missing the effect, if it is there, i.e. from the accepted beta error or, conversely, from the statistical power of the test I am targeting.

Because all this remains somewhat abstract without numbers, we will now illustrate the whole thing with a few examples.

Examples

1. *Aspirin and lipid-lowering drugs to prevent heart attacks*

A classic example is the so-called "Physicians' Health Study", which was conducted in the 1980s [5]. Physicians were invited to participate in a primary prevention study in which different preventive measures were blinded and tested over a longer period of time. Among other things, they investigated whether aspirin could prevent heart attacks. The idea behind this is simple: aspirin is a classic blood thinner and anti-inflammatory. The pathological processes that lead to heart attacks are inflammatory processes in the vessel wall and comparatively viscous blood, and the combination of both processes results in reduced blood flow to vital muscle areas in the heart. If one could both inhibit the inflammatory processes and improve the viscosity of the blood, the probability of heart attacks should be reduced in the group taking aspirin as a preventive measure - hence "primary prevention"; for the doctors were not ill. A small effect was expected, so 11,037 doctors received aspirin and 11,034 received placebo. The study was stopped after about 5 years by the monitoring board, i.e. a group that monitors the data during the study. This was because it had been shown that the aspirin group had significant advantages over the placebo group.
Fewer heart attacks had occurred in the aspirin group, namely 47% risk reduction (the relative risk was 0.53, which is how the 47% risk reduction is calculated). The alpha level of 5% was far below: the p-value was $p < 0.0001$ for heart attacks. So one made a mistake in less than 1 in 10,000 cases when claiming that aspirin can help prevent heart attacks. An astonished and frightened exclamation went through the specialist press and the lay press at that time. Suddenly it looked as if the whole world had to take aspirin for breakfast. Really? How big was the effect?

There were a total of 104 cases of heart attacks, 5 of which were fatal, in the aspirin group and 189 cases of heart attacks, 18 of which were fatal, in the control group. This only becomes really comprehensible when you see the absolute figures: 104 out of 11,037 people in the verum group had a heart attack, i.e. 9 per mille, and 189 out of 11,034 in the placebo group, i.e. 1.7%. If you put 9 per mille in relation to 1.7% (0.9/1.7), you get the relative risk of 0.529 or 0.53 given in the text. Compared to the control group, the aspirin group shows 47% fewer heart attacks.
However, since there were only 1.7% heart attacks overall, the actual absolute effect is very small.

One can also transform this dichotomous effect size into a continuous effect size d [6] and then obtain the equivalent of $d = 0.05$. If one considers that NICE, the English regulator, has set the clinical relevance limit for many therapeutic procedures, for example for the efficacy of antidepressants, at $d = 0.5$, then this effect size is tiny. Is it worth the effort? Apparently not. Because the effect, although small, comes at the price of side effects. For example, aspirin causes more brain hemorrhages, strokes and the like, which outweighs the positive effect of aspirin. For this reason, aspirin has fallen out of fashion again for the *primary* prevention of heart attacks, although it is still used for secondary prevention, i.e. in patients who have already had a heart attack.
This is a classic example of how a relatively small effect with a very large sample can still be "made" statistically significant with a small beta error, i.e. with a low probability of being missed, i.e. scientifically demonstrated to be present with a relatively low probability of error. The question that then arises is: How useful is this effect in a clinical sense?

Do enough people benefit sufficiently so that it is justifiable, in a social and clinical sense, to treat all those who will not benefit at all in the end? Or to put it another way: Is the ratio of profit and costs - financial costs and side effects - such that it makes sense to actually implement the scientifically proven result in practice? The aspirin example shows: there are quite a few cases where it is "scientifically proven that" something works or exists, in this case the reduction of heart attacks through the continuous preventive intake of aspirin, but it is still nonsense to use this finding practically.
Other examples for which the discussion is not quite over, although one could also argue about them, are the large primary prevention studies on the efficacy of lipid-lowering drugs, well summarized and critically discussed by Penston [7]. The large studies looked at between 4,100 and 19,300 patients over several years and examined the incidence of mortality, or heart attack, or stroke. The percentage of patients who were in the treatment group and had a success compared to the placebo group, i.e. none of the events studied, ranged from 1.4 to 3.8%, depending on the study and outcome parameter. In other words: between 96.2% and 98.8% of the patients were treated without the treatment being necessary or having a visible success, simply because the total number of patients in whom such an event as a stroke or heart attack occurred at all is very small. So if, as in the GISSI 3 study, 1.4% of more than 19,000 patients benefit, that is a very small effect. Let us assume for the sake of simplicity that out of

If 20,000 patients, half of whom receive lipid-lowering drugs and the other half placebo, have about 3% of a heart attack or stroke during the observation period, that would be 600 patients, i.e. an average of 300 per group. If we assume an RR = 0.5, i.e. a patient in the treatment group would have half the probability of having a heart attack under lipid-lowering therapy, then that would be 400 in the control group and 200 in the treatment group [8]. So that's twice as much and sounds like a lot. In absolute terms, however, the difference is very small: because 200 out of 20,000 is 1% of the sample and 400% is 2%, together 3% or 600. So you have to treat a total of 19,400 people with lipid-lowering drugs to get such an effect, who would not benefit in any way because they would not have had a heart attack. That is how these figures should be read. In other words, here a tiny little effect is made statistically significant with a huge expenditure of money, without any real account of how significant the effect is, whether it is proportionate to problems and side effects and the costs. Lipid-lowering drugs are cheap in single doses, but expensive in quantity. Moreover, they not only inhibit unwanted fat synthesis in the body, but also lower fat levels of essential fatty acids and the important coenzyme Q10, which is important for cell function, which is why they also very often cause very painful side effects such as muscle pain. Therefore, one could indeed argue whether the effect found is actually clinically useful, especially since a control of blood lipids can be achieved cheaply, without side effects and absolutely permanently through exercise and dietary changes.
We can illustrate this once again with a current example, the JUPITER study, one of the largest lipid-lowering primary prevention studies conducted to date, and specifically with the evaluation of the women's cohort, which turned out relatively well [9]. By the way, the whole study was heavily debated and some authors accused the authors of the JUPITER study of fraud [10], but we do not want to pursue this further now.

A total of 17,802 patients were included in this study, 6,801 of whom were women, the only ones we are concerned with here. They had a non-specific elevated C-reactive protein. This is a non-specific

inflammatory marker which served in this study to detect potentially at-risk individuals in the overall population so that the potential hit rate of the study would be higher. The absolute incidence rates for coronary heart disease in the included women with elevated C-reactive protein were 0.56 in the lipid-lowering group and 1.04 and in the placebo group, standardized to 100-person-years (this is done because the observation times are different for each). This means that if you observe 100 people for one year, then there are almost exactly half as many cases of coronary heart disease in the lipid-lowering group as in the placebo group.
This study was also terminated prematurely by the Trial Monitoring Board after 1.9 years because the treatment group had achieved statistically significant successes compared to placebo, i.e. the previously specified significance threshold of 5% had been reached, although it was initially thought that it would have to be observed for 5 years. The formal endpoint of the study, i.e. the target criterion, was a so-called composite endpoint, in which all cases of heart attack, stroke, need for hospitalization or surgery to restore blood flow or death were included. This was a total of 39 in the treatment group out of 3,426 women and 70 out of 3,375 women in the placebo group [11], i.e. slightly more than twice as many women in the control group suffered one of the predefined disease events.
As can be seen, the statement that an event occurred only half as often in the treatment group, or that the treatment group is twice as superior to the control group, only has real significance if it is specified by the total number of events. However, the incidence, i.e. the occurrence of the event is approximately only 1%, and twice as high in the control group, i.e. 2%. While the public presentation always operates with *relative* successes - "twice as good as in the control", "only half as many deaths, heart attacks, Strokes with lipid-lowering drugs" - this statement can only be evaluated if one knows how high the *absolute* number of occurrences was, a fact which, incidentally, is deliberately concealed in the publicly accessible abstract of the JUPITER study.

A power analysis is not necessary here, since the study was terminated prematurely. What is not superfluous is the consideration of whether it makes sense to treat a total of 6,800 others because of the 31 out of 70 women in the control group whose disease event could have been prevented, in whom none of the events would have occurred during the observation period. What happens when statins become a daily food because the whole world believes their primary preventive efficacy is scientifically proven? What side effects or long-term problems of a different kind do we trade ourselves? Maybe Alzheimer's because the balance of essential fatty acids shifts? Maybe chronic fatigue because cellular functions are reduced by the reduction of Q10, or chronic pain syndromes? Maybe completely different problems that we don't yet have in mind? We do not know. Because such questions are not investigated even by relatively long-lasting clinical studies and therefore cannot be answered.
As you can see, even tiny effect sizes can become significant with a lot of money, a lot of people and a long observation period, and with the corresponding statistical power, i.e. a high probability that the effect will become visible and a low risk that it will be overlooked. But is it useful? Do we want it? Do we want to pay for it? These questions are not answered by studies, but only by informed discourse.

2. *Manual therapy for complete cruciate ligament rupture*

The cruciate ligaments stabilize the knee. Sports injuries, such as skiing, often lead to the tearing of the cruciate ligaments, a very painful injury. Spontaneously, such injuries heal completely in perhaps two years, as we know from recent studies [12]. Therapeutically, however, usually surgery is performed.

This has become routine. However, it is an expensive routine and one that not infrequently leads to complications, such as infections or wound healing problems, or perhaps, after years, to secondary problems such as arthrosis. That's why top athletes have always looked for alternatives.
In Hallein near Salzburg there is a remarkable manual therapist, Mohammed Khalifa. Every day he devotes about 4 hours to preparing for his therapeutic work, through physical and mental exercise. He treats patients with complete ligament ruptures using only his hands. This is extremely painful, but very effective, so effective that many sports greats, from top footballers in the Bundesliga to tennis world champions and top climbers have had themselves treated by him. Because after the treatment, they are fit for action again within a few days, despite a complete rupture.

We wanted to investigate this and conducted a clinical study on 30 patients with complete cruciate ligament rupture [12, online at http://dx.doi.org/10.1155/2014/462840]. Patients were randomized to receive either Khalifa therapy or conservative but very good physiotherapy. The target criterion was a knee assessment value recorded by the physician, the so-called knee score.
IKDC value; this captures the objectively assessable functional capacity of the knee. In addition, we verified the therapy effects by imaging, which could show whether or not the ligament had healed together after 3 months. We measured the success of the therapy on the day after treatment and 3 months later. At this time, all patients also received an MRI again for imaging. After 3 months, the cruciate ligament had healed in 7 of the 15 patients in the treatment group and in none of the patients in the control group. The main outcome measure, the IKDC score, showed very strong and significant effects already on the day after treatment and 3 months later. Was the effect not only statistically significant but also clinically meaningful? Was it a chance finding or what was the statistical power? The graph and results table shown in the original paper provide information about this, and those who have been paying attention can also follow the analyses themselves. The effect size immediately after treatment isquite large with d = 1.77 and still very large after 3 months with d = 1.19. The treatment group is therefore more than one standard deviation above the control group. The treatment group is thus more than one standard deviation above the control group, after a single treatment. The statistical power of the analysis was 0.88, which is quite good, even with only 15 patients per group. Now is this an effect that is significant? We did not compare against OP. We estimate that the same functional improvement would be seen against surgery in the long term, but probably not in the short term, because the post-operative complaints would initially limit the functionality of the knee more. The surgery costs more money and carries more risks. The risk of the Khalifa treatment is the one-time pain and a risk, not yet explored in long-term studies, that the therapy will be unsuccessful for some people. So here, too, one has to consider the advantages and disadvantages and assess the effect size on the basis of the available alternatives, their advantages and disadvantages.

As you can see, even with a comparatively small study of a total of 30 participants, an effect can be made visible well if it is large enough.

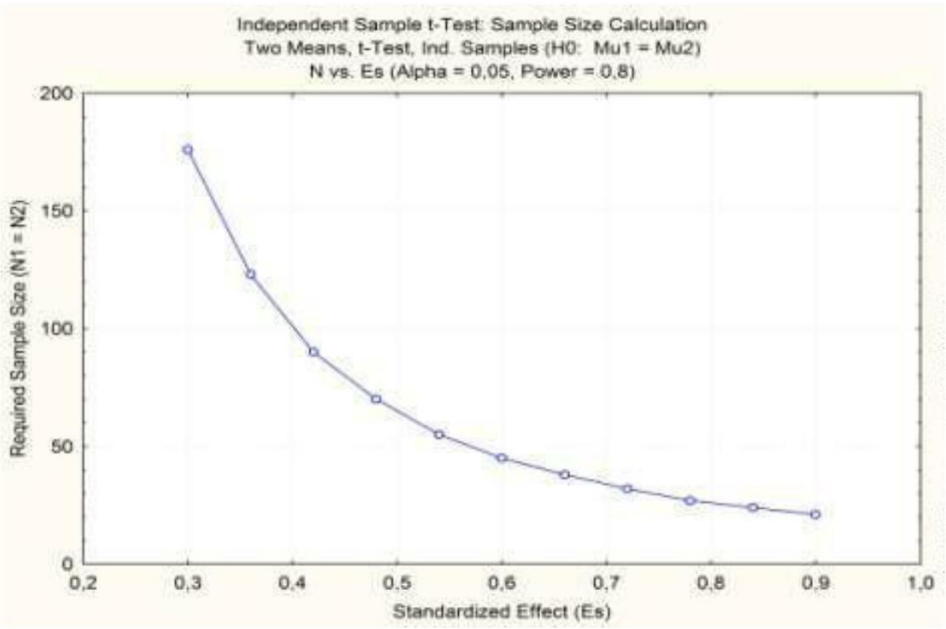
Own exercises:
There is a nice free programme, G-Power. With it, anyone can do their own power analyses and understand what I have said here http://www.gpower.hhu.de/

Here are two graphs that illustrate how effect size, sample size and power correlated. I created them with my STATISTICA programme, but G-Power also offers the possibility to create such graphs.

It is very easy to see that if you aim for a reasonable statistical power, 0.8 is usually considered useful and the minimum, and you thus have a chance of detecting an effect in 80% of all cases, then depending on the effect size present you will need different sample sizes.

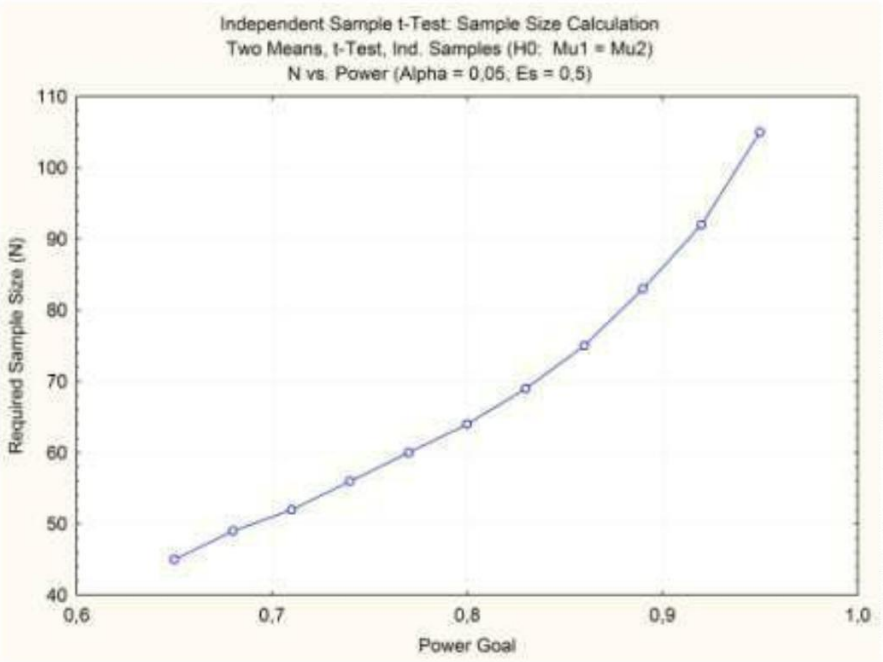
For the effect we found of more than one standard deviation, the power was very good even for our small study. Studies that are done frequently, with about 30 people per group, only have a chance of finding the effect with decent power if it is larger than d = 0.75, as you can see from the graph. These are about the kind of effects one might expect from a good psychotherapy. For comparison, the adjusted effect size of antidepressants is d = 0.38 [13], about the same as estimated by a very conservative meta-analysis for the treatment of anxiety or depression by meditation [14]. Medium effects of about d = 0.5 or smaller require at least 70 people per group and if the effects are small, about d = 0.3, you need to recruit 180 per group. Conversely, one can conclude: if someone needs 5,000 people per group, then we are dealing with tiny effects.

Fig. 1 - The relationship between sample size and effect size with an assumed power of 0.8



Similarly, Figure 2 shows that if you want to find an effect of half a standard deviation, i.e. d = 0.5, which is assumed to be fixed here - this is the effect that NICE postulates as clinically relevant in depression treatment - then you have to recruit more and more subjects with increasing power or less willingness to overlook an effect.

Fig. 2 - The relationship between power and sample size with an assumed effect size of d = 0.5



We ourselves have also been confronted with this problem. We have conducted what is probably the largest and perhaps the first active controlled trial of mindfulness meditation for the treatment of fibromyalgia, a three-arm study in which we compared MBSR, a form of group mindfulness training, with an active control condition, stretching and relaxation, and with a waiting group [15]. We assumed an approximate effect size of d = 0.55, based on our earlier meta-analysis [16]. Stupidly, we underestimated the strength of active control. Although the mindfulness condition offered some advantage, as was especially evident in the interviews, we could not statistically validate the effect against active control. This is because the effect size was only about d = 0.3 and from Figure 1 you can see that you need about 180 patients *per group* to ensure an effect of this size, so our power was too small. That was about the number of patients we had for the whole study. So is mindfulness useless for fibromyalgia? Probably not. Because even an effect size of only one third of a standard deviation is a success in this disease. Because fibromyalgia is difficult to impossible to treat and often a lifelong burden of continuous pain for patients. Have we shown that mindfulness can be considered for this as a treatment strategy? Not in the strict sense, because our study was not significant. A larger study could now try to make up for that. To do that, someone would have to raise much more money than we have available; unlikely, given the data. Or someone could accumulate the effect sizes across different studies.

That would be the goal and potential outcome of a meta-analysis, and that will occupy us in one of the next blogs.

[1] Fisher, R. A. (1971 (orig. 1935)). The Design of Experiments. New York: Hafner.
[2] a small numerical example is given by the following table:

	treats	untreated
Success (a)	40	20
Failure (a')	20	40
Total number	60	60

RR = a/ n1 : a'/n2 = 40/60 : 20/60 = 40:20 = 2; means: twice as many patients in the treated group were successful. Note: the total number, i.e. the respective denominator of the ratios, can be ignored if the group size is the same, because the numbers are then reduced. Therefore, one can simply calculate 40:20.
[3] Which mean value one sets first, whether that of the control group or that of the treatment group, is a matter of definition and taste and also depends on the polarity of the measures used. As a rule, the equation is formulated in such a way that a positive value for d defines an effect size in the sense of treatment. For example, if one records depression values or blood pressure, where a high value denotes a lot of depression or high blood pressure, then one must subtract the value of the treatment group from that of the control group. The reverse is true for quality-of-life scales, which are usually constructed in such a way that a high value is desired. One has to make sure that this is done consistently for all the values examined.
The standard deviation that you put in the denominator is usually the averaged, or "pooled" standard deviation of the two groups. You can also use the larger standard deviation if you want to estimate robustly. This is because the larger the standard deviation, the smaller the d.

[4] The basics of power analysis were elaborated by Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences. Hillsdale: Lawrence Erlbaum. (Original published 1977: New York: Academic Press).
[5] Steering Committee of the Physicians' Health Study Group (1988). Preliminary report: findings from the Aspirin component of the ongoing physican's health study. The New England Journal of Medicine, 18, 262-264.
[6] This is a procedure with which you can make studies that are based on different metrics approximately comparable. It can be used when conducting a meta-analysis, for example. Relevant manuals such as Rosenthal, R. (1991).
Meta-Analytic Procedures for Social Research. Newbury Park: Sage contain the relevant information. In this case, one takes an arcsine transformation. It transforms rate ratios into d-values, at least approximately. The same is achieved by the formula of Hasselblad, V., & Hedges, L. (1995). Meta-analysis of screening and diagnostic tests. Psychological Bulletin, 117, 167-178.
[7] Penston, J. (2003). Fiction and Fantasy in Medical Research: The Large-Scale Randomised Trial. London: The London Press.
[8] We see from formula [2] for the rate ratio that it is in principle independent of the total number. If the total numbers in both groups are exactly the same, then they cancel each other out and what remains is the absolute ratio of successes in one group to successes in the other. Only if the groups are of different sizes does the formula adjust for the proportional differences.
However, this also means that one must always keep an eye on the occurrence of the event in the population in order to be able to estimate the effect.

[9] Mora, S., Glynn, R. J., Hsia, J.,MacFadyen, J. G., Genest, J., & Ridker, P.M.(210).Statins for the primary prevention of cardiovascular events in women with elevated high-sensitive c-reactive protein or dyslipidemia - Results from the justification for the use of statins in prevention: An intervention trial evaluating rosuvastatin (JUPITER) and meta-analysis of women from primary prevention trials. Circulation, 121, 1069-1077.
[10] de Lorgeril, M., Salen, P., Abramson, J., Dodin, S., Hamazaki, T., Kostucki,W.,et al. (2010). Cholesterol lowering, cardiovascular diseases, and the rusvastatin-JUPITER controversy: A critical reappraisal. Archives of Internal Medicine, 170, 1032-1036.
[11] Because the groups are approximately the same size, the group size can be neglected for the frequencyofoccurencea n d HR (hazard ratio) = 39/70 = 0.56 is calculated.
[12] Further literature, also on the measure used and spontaneous healing rates in our originalstudy:Ofner, M., Kastner, A., Wallenboeck, E., Pehn, R., Schneider, F., Groell, R., et al. (2014). Manual Khalifa therapy improves functional and morphological outcome of patients with anterior cruciate ligament rupture in the knee: A randomized controlled trial. Evidence Based Complementary and Alternative Medicine, Art ID 62840. https://doi.org/10.1155/2014/462840
[13] Turner, E. H., Matthews, A. M., Linardatos, E., Tell, R. A., & Rosenthal, R.(2008). Selective publication of antidepressant trials and its influence on apparent efficacy. New England Journal of Medicine, 358, 252-260.
[14] Goyal, M., Singh, S., Sibinga, E. M., Gould, N. F., Rowland-Seymour, A., Sharma,R . , et al. (2014). Meditation programs for psychological stress and well-being: A systematic review and meta-analysis. Journal of the American Medical Association - Internal Medicine, doi: https://doi.org/10.1001/jamainternmed.2013.13018.
[15] Schmidt, S., Grossman, P., Schwarzer, B., Jena, S., Naumann, J., & Walach, H.(2011). Treating fibromyalgia with mindfulness-based stress reduction: results from a 3-armed randomized controlled trial. Pain, 152, 361-369
[16] Grossman, P., Schmidt, S., Niemann, L., & Walach, H. (2004). Mindfulnessbased stress reduction and health: A meta-analysis. Journal of Psychosomatic Research, 37, 35-43.