POWER, MAGNITUDE OF EFFECT AND SAMPLE SIZE

Allan Wilson MD PhD

Power, magnitude of treatment effect (effect size) and sample size interact to determine the signal-to-noise ratio (***S/N***) of a clinical trial. The object is to determine if the treatment groups differ significantly from each other. For many trials this is decided by a ***t-test*** or an ***F-test*** (analysis of variance) but other statistics are also used. An ***F-test*** compares treatment groups by looking at the ratio of the variability between the groups (signal) to the variability within the groups (noise). (For a two-group design, a ***t-test*** is equivalent to an ***F-test***, with $F = t^2$.)

The significance of the differences between treatment groups is determined by ***p*** (or ***alpha level***) - the probability that a real treatment difference would be detected. Of course one might conclude erroneously that there is a difference between treatment groups when there is not. This is a Type I error. Conversely, one might miss an actual difference – a Type II error. ***p*** is generally set at .05 for clinical trials but other (usually more stringent) criteria such as .01 or .005 are sometimes used. A ***p*** of .05 means that five times out of 100 a false positive error will occur.

POWER

Power is the ability of a study to have the signal stand out from the noise. It increases with sample size ***n*** because a larger ***n*** means decision-making is based on more information, increasing the accuracy of any conclusions**.** Of course data are acquired at considerable cost. Thus, clinical trials are designed to optimize the relationship between power and ***n***, with power of about 0.8 generally being considered the best compromise. (There are cases where more or less power is used.)

Power calculations are somewhat complicated and take into account the type of statistical test to be used (eg: ***t-test***), the desired ***p*** value (the smaller the ***p*** the more convincing the demonstration of any difference between treatment groups), the ***n*** under consideration (***n*** may be constrained, for example, by the availability of funds or subjects), and the magnitude of treatment effect. Whereas the relationship between ***p*** and ***t*** or ***F*** is somewhat inverse (although ***p*** is not proportional to ***1/F or 1/t***), magnitude of effect is more complicated.

MAGNITUDE OF EFFECT (EFFECT SIZE)

This parameter is determined from previous experience with the treatment(s) of interest. Effect size refers to the magnitude of the difference between treatment groups. It is somewhat subjective at best and can be very subjective for novel treatments. The investigator determines what difference between treatment groups would be convincing (ie: make sense clinically) regardless of the ***F*** or ***p*** values. This can be important in situations where the ***S/N*** is large. For example, two small groups of 21-year old males are studied for temperature changes following administration of a metabolism-altering medication. One group averages 37.1^{o}C and the other 37.3^{o}C. If the groups were highly similar in their background characteristics, a ***t-test*** might result in ***p* < .05**, indicating a significant difference. However, the investigating endocrinologist might not accept that this is a clinically meaningful difference based on previous experience with temperature, similar drugs and patients. She might conclude there is no clinical difference in temperature despite the significance level of the ***t-test***.

Magnitude of effect can be viewed as given a supposed background noise level, and can be written as a multiple of the background noise $k\sigma$, where σ is the standard deviation. To complicate matters, magnitude of effect must be specified before starting the trial, posing a problem for studies involving novel drugs or treatment populations with which there may be little or no clinical experience. There are numerous ways of estimating magnitude of effect and, in the context of clinical trials, all rely on previous clinical experience.

POWER ANALYSIS

The power analysis takes these parameters into consideration to optimize the chances of the study design giving an accurate assessment of treatment effectiveness (***S/N***). Statistical programs allow the investigator to manipulate the parameters to achieve the best balance. Generally, **alpha** (***p***) is set at 0.05; acceptable power at 0.8, magnitude of effect between 0.5 and 0.7 (or, technically, 0.5σ where σ is assumed to be 1 for simplicity***)*** and the sample size ***n*** is computed.

BACKGROUND TO THE ROI WIDGET

IMC's ROI (Return on Investment) widget is designed to give users or potential users of electronic compliance monitoring (ECM) a feel for the return on investment. Although ECM comes at a cost, the return on investment typically dwarfs the initial outlay. Many factors, some of which are likely to be unknown, underlie any ROI calculation, and the ROI widget makes a number of assumptions. Thus, the results should be viewed as estimates, but the ROI calculation is based on conservative assumptions and should, in most cases, underestimate the true ROI.

The ECM quantifies patient non-compliance, a source of noise, and reduces its contribution to the denominator of the ***S/N*** ratio. This allows the signal to stand out better. This is what generates the ROI.

COMPLIANCE, POWER AND THE *F* STATISTIC

In the simplest case we have two treatment groups of equal size ***n*** and equal average compliance ***C***. We also make the plausible assumption that the compliance effect is linear as this is the most conservative approach. The effect of compliance may be non linear, which would increase the ROI estimation (except for the improbable case of $r < 1$, in which the ROI would be underestimated).

(1) $$F = (\text{average compliance proportion})^2 * (F_{\text{under perfect compliance}})$$

If we assume the effect of compliance is the r^{th} power,

(2) $$F = (\text{average compliance proportion}^r)^2 * (F_{\text{under perfect compliance}})$$

Thus, improving compliance by ***x*** percent will increase ***F*** by about **2x** percent assuming a linear compliance effect (***r*** =1). If the realtionship were really a higher order one, ***F*** would be increased significantly more.

For group treatment effects μ_1 and μ_2, ***F*** under linear compliance effect will be:

(3) $$F \propto C^2 (\mu_1 - \mu_2)$$

The relationship between power and compliance for ***n*** = 25, 50, 100 and 200 is shown in Figure 1.

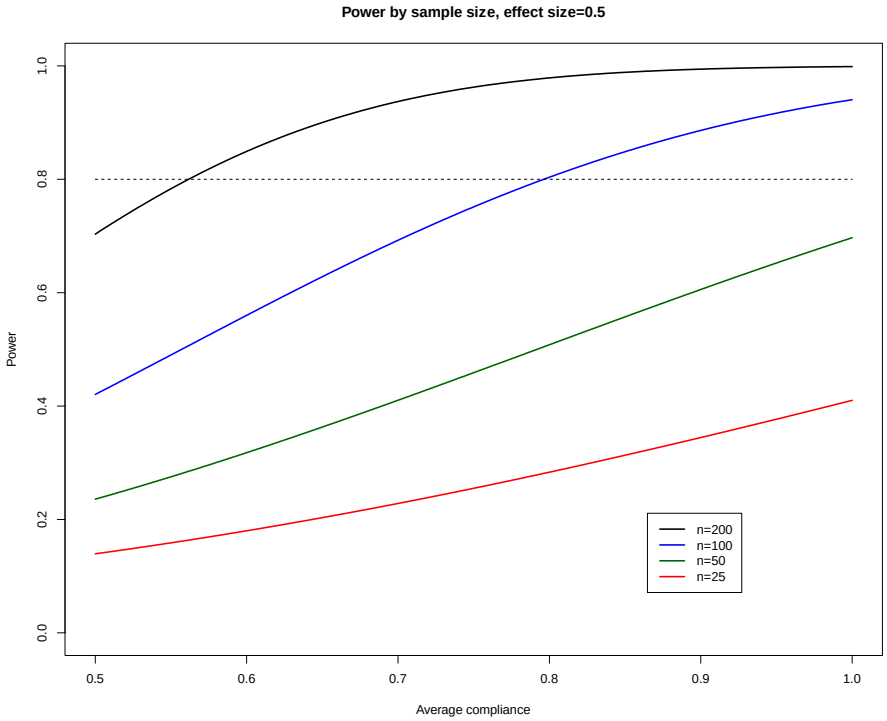

Fig. 1. Power versus compliance for magnitude of effect = 0.5

Note that for ***n*** = 100 (blue) the study would be underpowered according to the widely accepted power criterion of 80 percent if average compliance were 70 percent, but would be acceptably powered for a clinical trial (.80) at 80 to 85 percent average compliance, and power continues to increase as a function of higher compliance levels.

POWER AND MAGNITUDE OF EFFECT

The relationship between power and magnitude of effect ($\mu_1 - \mu_2$) is also helpful in understanding the ROI widget. Figure 2 shows the relationship between power and compliance for the commonly used effect size of 0.5 and Figure 3 for the considerably more stringent effect size 0.7. Estimated magnitude of effect for clinical trials generally ranges from 0.5 to 0.7.

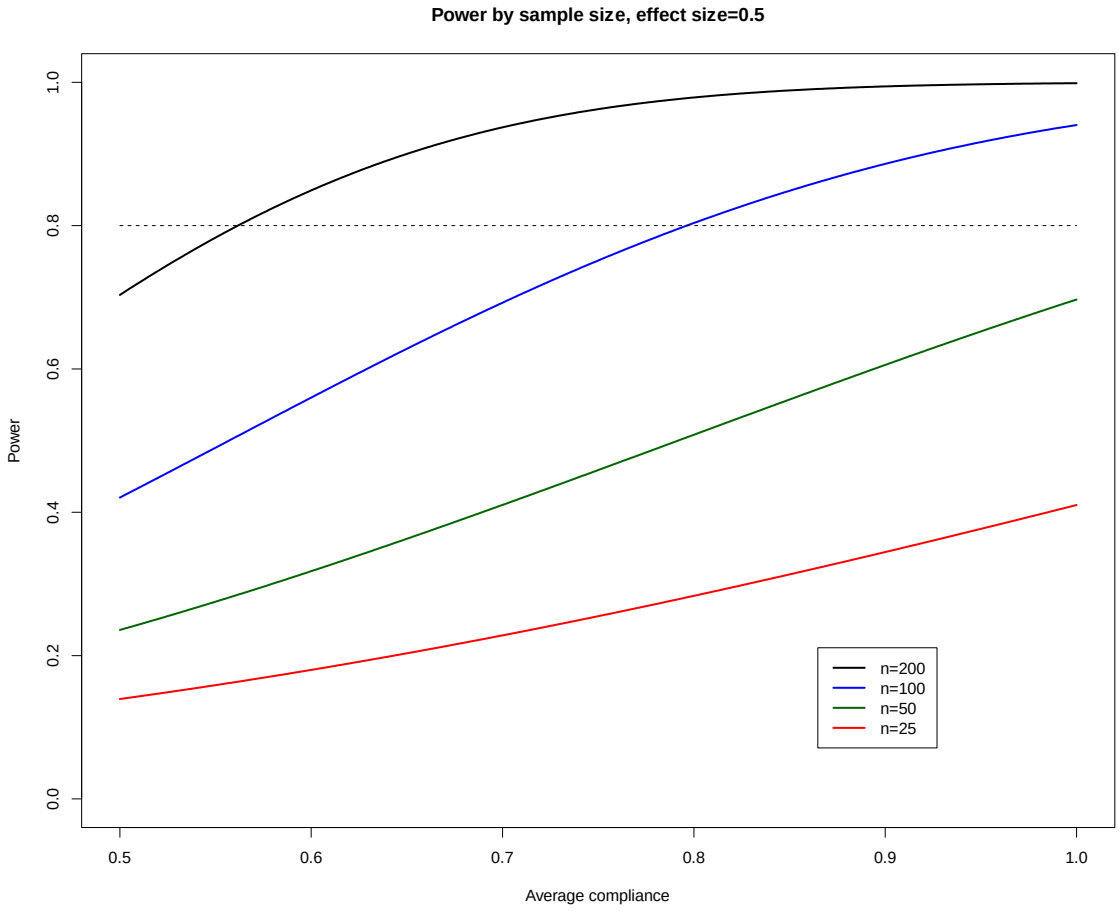

Fig. 2. Power versus compliance for magnitude of effect = 0.5; *n* = 25, 50, 100 and 200

Figure 2 shows that for a trial with a sample size of 100 (blue) and magnitude of effect = 0.5, acceptable power (0.8) is achieved only if average compliance is greater than 80 percent.

Figure 3 shows the power-compliance relationship for magnitude of effect = 0.7, an estimate that would be unrealistically high for most clinical trials. Using this stringent criterion, acceptable power of 80 percent would be achieved for average compliance of 0.8 using a sample size of only 50 (green).

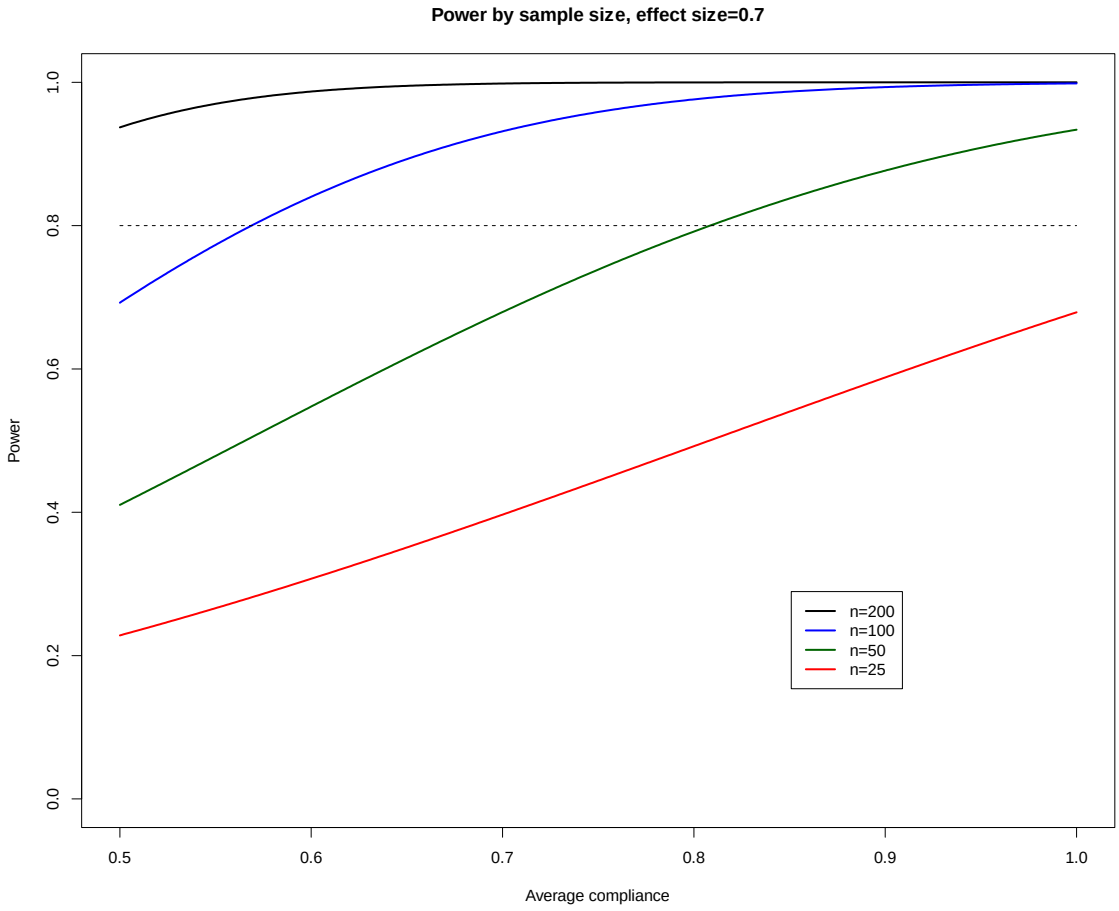

Fig. 3. Power versus compliance for magnitude of effect = 0.7; *n* = 25, 50, 100 and 200

SAMPLE SIZE

Figure 4 shows what compliance monitoring is all about.

For a given treatment effect $\mu_1 - \mu_2$, n is proportional to the reciprocal of the square of the compliance:

(4) $$n \propto \frac{1}{c^2 (\mu 1 - \mu 2)^2}$$

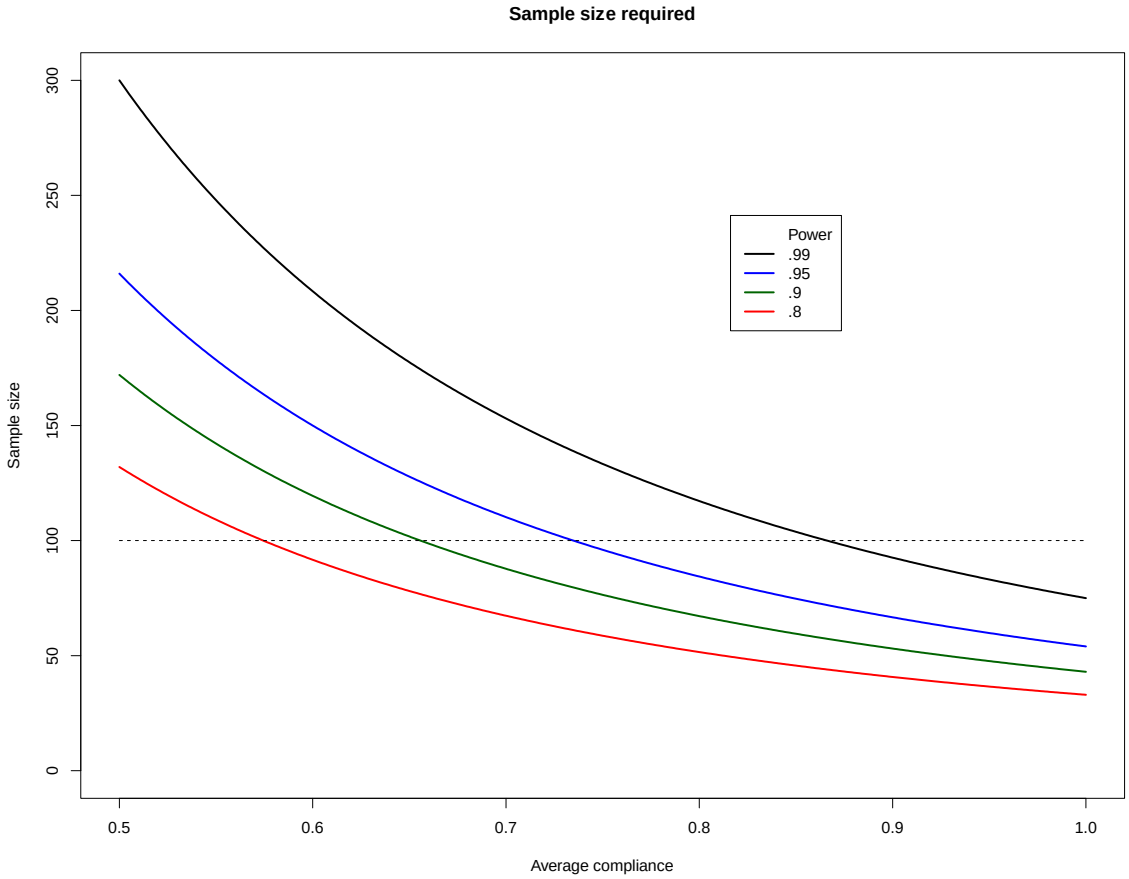

Fig. 4. Power, compliance and sample size

As can be seen in Figure 4, for an acceptable 80 percent power level (red curve), increasing average compliance from an extremely poor 50 percent to a still low 70 percent would reduce the required sample size by 48 percent. Increasing compliance to 90 percent would reduce the required ***n*** by 69 percent.

As for the ***F*** statistic, improving compliance by ***x*** percent is equivalent to increasing the sample size by about ***2x*** percent, assuming the most conservative case of linearity. As can been seen from the above figures, the relationship tends to non-linearity. For a pure quadratic effect an ***x*** percent increase in compliance would be equivalent to a ***4x*** percent increase in ***F***. In the real world, a combination of linear and quadratic effects is most likely, yielding somewhere between ***2x*** and ***4x*** percent improvement with compliance. The ROI widget takes the more conservative ***2x*** percent stance.

COMPLIANCE DATA IN PRACTICE

From a strategic perspective there are two approaches to using the power-enhancing effect of compliance data. In most applications the information will be used to reduce the sample size consistent with maintaining the desired power level, magnitude of effect, and ***p.*** However in some situations it may be desirable to make a compelling case for a drug's efficacy by demonstrating a large ***F*** and correspondingly small ***p.*** Here, the sample size suggested by the initial power calculations can be maintained and the compliance data can be reflected in a larger ***F*** and smaller ***p*** to increase confidence in the results.