POWER & SAMPLE SIZE ESTIMATES:

Welcome back to our evidence anxiety series where we've been discussing approaches to appraise truth claims. In the previous episode we talked about the importance of the comparison group. In this episode we'll talk about the importance of considering study power and sample size estimates. As a point of clarity, samples can be things or observations that the researcher measures or counts to create data for analysis. When we talk about the size of a study it's actually in reference to the amount of data collected and it's important to collect enough to be able to detect an effect, or if it's an observational study, to show a relationship. To keep things simple we'll just talk in terms of effects from here on.

Study power and sample size estimates go hand-in-hand and the terms are often used interchangeably because the study size impacts the power of a study, but technically, they are not the same thing. By definition, power is the probability of detecting a true difference, whereas the sample size estimate is the calculated number of study samples or observations, or simply, how much data are needed for detecting a difference[1].

If researchers claim there's a difference in effects between comparison groups where none actually exists, then they have made a type I error, and have falsely reported that the treatment has an effect. If the researchers report no difference between comparison groups where a true difference actually does exist then they have made a type II error, or have falsely reported 'no treatment effect'. It's the type II errors that are impacted by the sample size and study power.

As a rule of thumb, if the researchers expect the treatment effect to be small, then the study has to be big, conversely, the smaller the study, the more likely it is that a small or rare effect is missed[2]. This relationship is the same regardless of the treatment applied, be it a surgical method, a behavior modification strategy, a drug, or a vaccine. Put even more simply, big studies have more power to detect differences.

Here's where it gets tricky. It's important for the reader to take note of how the researchers interpret study findings if they report that no treatment effect is found. In statistical terms the researchers would say that "the effect is not statistically significant". In some ways, research studies are similar to a legal case in that a verdict of "not-guilty" shouldn't be interpreted as "innocent"[3] just as a statistical finding of "no treatment difference" shouldn't be interpreted as "the treatment has no effect". Both legal cases and research studies are not designed to prove innocence, or in the case of research, no effect. It's more accurate to report that the evidence was insufficient in the study to conclude there was an effect. So if you are reading a study where the authors conclude there was no effect, expect also for them to explain how they decided on the size of their study to give you an idea of how much evidence that was collected before making that conclusion. This is of particular importance if a study is underpowered and fails to detect an adverse outcome.

In a practical sense the size of a study may be limited by available resources and it's important for reader to know if this also constrained the power of the study.

Study size estimates are calculated from four basic components: The first is the desired study power and the second is the accepted probability of making a type-I error. The researcher selects both at the onset of study planning. The third component is the smallest clinically meaningful effect that the researchers wish to detect, and the fourth is a measure of the expected consistency of the effect in the general population.

By convention the power is usually set at 80 %, the type-I error probability is set at 5% and the researchers generally have an idea of reasonable estimates for the latter two. Often the calculation is straight forward and can be done manually but in the case of complex study designs involving different levels of group comparisons, or studies involving repeated sampling, the formula may be fairly complex. In such cases it is not unusual for researchers to seek statistical expertise to help with the calculation[1].

For researchers, calculating an appropriate study size estimate is important to avoid making claims that may cause harm or waste valued public trust and resources. When the researcher spells out clearly how they decided on the size of the study it helps the readers to determine if the

findings are meaningful. A big give-away for the reader indicating they should be cautious about a study is that the researcher's confidence in the findings are not presented in statistical terms and also that there is no accompanying discussion of study power or of sample size estimation. In an upcoming infographic we'll discuss how to know if a statistically meaningful effect is also clinically meaningful.

References: (the Feinberg paper is a terrific read)

1. Jones SR, Carley S, Harrison M. An introduction to power and sample size estimation. *Emerg Med J*. 2003;20(5):453-458. doi:10.1136/emj.20.5.453
2. Wilson Van Voorhis CR, Morgan BL. Understanding Power and Rules of Thumb for Determining Sample Sizes. *Tutor Quant Methods Psychol*. 2007;3(2):43-50. doi:10.20982/tqmp.03.2.p043
3. Feinberg WE. Teaching the Type I and Type II Errors : The Judicial Process. *Am Stat*. 1971;25(3):30-32.