Powering Clinical Trials: Achieving Relevance

Enrolling insufficient numbers of subjects to power a clinical study is often cited as a major reason for failure to achieve statistically significant endpoints, frequently expressed as a large *'p'* value (>0.05) or high probability or relevance for data. Simply using greater numbers of subjects (increasing 'n') to overcome inherent weaknesses in study design may be subject to criticism for ignoring potentially confounding factors common to such models. Some of these factors are discussed in more detail below.

Pre-study Odds

The chances of finding data that support a given hypothesis should remain relatively constant in a static or stable population. Failing to take account of pre-study variation in odds can introduce bias into results and increase the probability of error, as has been discussed by John Ioannidis, Professor of Statistics, Stanford University School of Humanities and Sciences, who stated:

"As has been shown previously, the probability that a research finding is indeed true depends on the prior probability of it being true (before doing the study) … instead of chasing statistical significance, we should improve our understanding of the range of R values (the pre-study odds) where research efforts operate. Before running an experiment, investigators should consider what they believe the chances are that they are testing a true rather than a non-true relationship."

This means that taking time to understand what a normal distribution looks like within the study environment allows for better evaluation of likelihood of identifying relevant data, the effects of chance and the meaningfulness of outliers.

A simple comparative figure shows us that as 'R' falls, the positive prediction of any observation is also reduced:

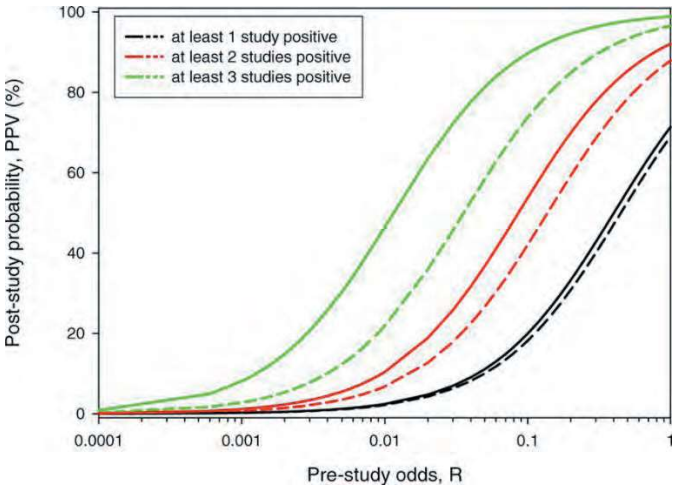

Figure 1. Probability of a true relationship when at least one, two, or three (out of ten) studies have statistically significant results as a function of the pre-study odds (r) of a true relationship (α = 0.05) Dashed lines refer to power of 0.2 (20%) and solid lines to power of 0.8 (80%) https://doi.org/10.1371/journal.pmed.0040028.g001

The Convention of 95% (Whereby Observed Events are 'True' if Powering is Sufficient)

The use of a 5% 'chance' factor does not adequately represent odds in populations with a high bias or with an unusual distribution of variability. A 5% measure for a random observation of falsely supportive, or contrary evidence may be adversely affected where evidence is either extremely sparse (e.g. seasonal) or has a low level of detectability due to little variance from normal (e.g. mildly deranged pathological values). The very fact that the distribution of events, or data observed relating to those events, may be represented by a Poisson distribution curve shows us that the more common the event, the more likely we are to observe an outlier. Thus, predictive values may be affected by both unexpectedly low and high rates of occurrence. Peter Bacchetti, Professor Epidemiology and Biostatistics at the UCSF School of Medicine, observes:

"…the positive predictive value (PPV) of p < 0.05 is an unacceptably poor measure of the evidence that a study provides. The fact of diminishing marginal returns precludes any meaningful definition of 'adequately powered' versus 'underpowered'; the goal of 80% power is only an arbitrary convention."

Effectively, the results obtained by a study should be predicated on the likelihood of outcomes and powered accordingly, rather than adding numbers until a p-value is seen to be acceptable. But does this mean a major rethink of clinical trial modelling, or is the industry already aware of such pitfalls?

Subject numbers are often based upon assumptions of drug or vaccine efficacy taken from previous trials and are susceptible to variations in cohorts or inclusion / exclusion criteria, sampling timepoints, et al. However, as a model becomes more frequently adopted over time, the value of data may increase, as both 'R' and 'n' increase in proportion to knowledge obtained as to pre-study probability and relative study size (e.g. meta-analysis is enabled).

The Null Hypothesis Assumes there is 'No Effect'

The idea that there is a simple positive or negative effect of an intervention is only belied by the variance in responses frequently observed. An intervention may have only a mild effect or work concomitantly with other host factors while showing a delayed action or providing downstream benefits. Powering only for black and white can ignore grey benefits or deficits – knowing what the likelihood of the dynamic range of effects is again essential to correctly understand cohort size requirements.

All of the above factors should be considered in the justification provided to regulatory authorities as part of the clinical trial authorisation application (ICH Topic E 9, Statistical Principal for Clinical Trials Step 4, Consensus guideline, 05Feb1998 Note for Guidance on Statistical Principles for Clinical Trial. Available via http://www.ich.org). Knowing your cohort variance is as essential as knowing your drug characteristics.

However, understanding the cohort only reduces the chances and cannot stop errors from entering the database. Variance may be

seen in many measurements which may often go un-noticed; these add to the 'noise' that must be considered by statisticians:

- Measurements regarding observational occurrences or changes that are subject to bias, e.g. constitutional symptoms such as fatigue, malaise, loss of appetite; or specific sensory experiences like pain, photophobia, parageusia;
- Poorly validated or novel ordinal scales or ranking (these are often subjective with little objective evidence);
- Unplanned variances in scheduled procedures, e.g. the timing and performance of collection, storage and testing of clinical specimens;
- Variance in assays (inter / intra-assay performance);
- Time (t) – at what threshold of time is a change relevant?
- Delta – how much of a change is needed for relevance?

Clinical trials are heavily regulated to ensure compliance to published guidelines, not only to protect subject safety and rights, but also to ensure that data has relevance and is reproducible. Bias may creep in even where knowledge of systems and structures are profound. According to a 2016 article by Monya Baker in *Nature*:

'More than 70% of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments. Those are some of the telling figures that emerged from Nature's survey of 1,576 researchers who took a brief online questionnaire on reproducibility in research.... More than 60% of respondents said that each of two factors –pressure to publish and selective reporting — always or often contributed. More than half pointed to insufficient replication in the lab, poor oversight or low statistical power. A smaller proportion pointed to obstacles such as variability in reagents or the use of specialized techniques that are difficult to repeat.'

[source: https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970]

A common misconception is that a clinical trial measures the probability of a null hypothesis being true given the observed data. This is, in fact, the wrong way around; estimations of probability (p) actually provide a measurement regarding the probability of observing the data given that the null hypothesis is true. It is very different to pronounce something as 'not guilty' (or not likely to be guilty) rather than innocent.

To ensure reproducibility and relevance of efficacy data observed in exploratory clinical trials, study conditions should be controlled and standardised as far as possible. Clearly translating data from the clinic to the community is more likely to be predictive when such data is statistically validated than potentially confounded by unaccounted errors or bias.

With the growth and standardisation of the Controlled Human Infection Model (CHIM), the potential for human challenge to provide statistically relevant Phase IIa/b data regarding efficacy is expanding. CHIMs offer a number of advantages over community trials for offering early efficacy and safety data:

- Challenge trials (CHIMs) have simple, quantifiable and measurable primary endpoints

- CHIMs offer reduced noise by controlling the environment and reducing complexities of individuals, infection and disease
- R is known and characterised (FIH studies)
- Outliers are eliminated or reduced
- Powering calculations are simplified; cohort sizes are reduced in relation to increases in PPV

Challenge trials may offer a route to diminishing the rate of late-phase (III) failures. With such studies now encompassing viral, bacterial and even parasitic challenge agents, it is hoped that, by providing statistically valid data, the design and performance of later, large-cohort clinical trials may be improved, leading to shortened pipeline development timelines and reduced capital burn. New consortiums such as the UK Human Infection Challenge in Vaccines (HIC-Vac) and workshops hosted by the International Association of Biological Standards (IABS) point to a renewed interest in CHIMs and a desire to standardise the model. As data becomes more comparable, meta-analyses may offer increased powering and new insights into infectious disease.

Adrian Wildfire

He has 30 years' experience in communicable diseases. He obtained his Fellowship in Medical Microbiology in 1990 and a Masters in Parasitology in 1998. He is the author/co-author of numerous papers with many of the UK's leading infectious disease experts and KoLs in tuberculosis, HIV, and influenza. He trained at the Wolfson Institute and LSHTM before moving on to lead teams within various institutions."

Email: adrian.wildfire@sgs.com