...the problem with powering

Adrian Wildfire
World Vaccine Congress
October 2017

WHEN YOU NEED TO BE SURE

THE PROBLEM WITH THE PROBLEM

Why Most Published Research Findings Are False

John P. A. Ioannidis

“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.”

Small sample size is not the real problem

Peter Bacchetti

“…the positive predictive value of $p < 0.05$ (PPV) 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.”

SOME SIMPLE FACTS ABOUT NUMBERS

1. Subject numbers (n) are often based on previous experience in similar trials
2. Most regulatory agencies now require a justification of sample size
3. A study with too much power may be costly and may claim significant results that are not clinically relevant
4. Any study that lacks power will not be significant – even if results are clinically meaningful*
5. Studies should have sufficient statistical power (>80%; preferably ≥95%;) to detect clinically meaningful differences between groups
6. A sample size calculation plus type of analysis should be considered early in the planning stages; n, one / two-tailed, CI etc.

BUT: PPV ignores distinctions between different p values below 0.05, such as $p = 0.049$ versus $p < 0.0001$

*There is a known publication bias against studies with negative findings

p PROBLEMS

Important:

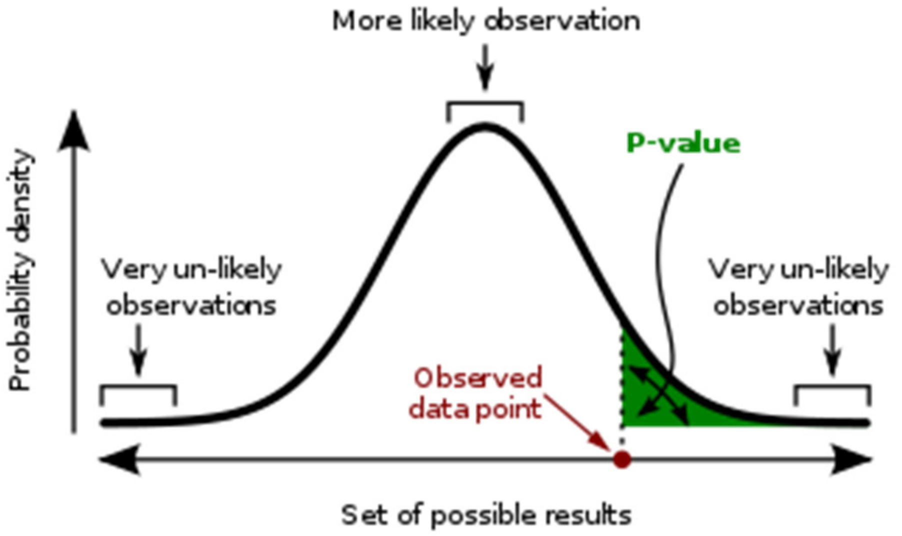
Pr (observation | hypothesis) ≠ Pr (hypothesis | observation)

The probability of observing a result given that some hypothesis is true is *not equivalent* to the probability that a hypothesis is true given that some result has been observed.

Using the p-value as a "score" is committing an egregious logical error: **the transposed conditional fallacy.**

	DECISION	
TRUTH	Accept H_0:	Reject H_0:
H_0 is true:	**correct decision P** *1-alpha*	**type I error P** *alpha (significance)*
H_0 is false:	**type II error P** *beta*	**correct decision P** *1-beta (power)*

H_0 = null hypothesis
P = probability



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

WHY DOESN'T n = TRUE / FALSE?

To see how powering affects predictive values we can observe how PPV falls in relation to the power 'n' of the study:

$$PPV = \frac{Power * R}{Power * R + 0.05}$$

Suppose you are in a field where 1 in 5 hypotheses is correct. R = ¼ = 0.25.

2.5 / 5 = correct

Power = 20% PPV = 0.2 * 0.25 / (0.2 * 0.25 + 0.05) = 0.50

4 / 5 = correct

Power = 80% PPV = 0.8 * 0.25 / (0.8 * 0.25 + 0.05) = 0.80

n = number
R = pre-study odds

TYPE 1 AND TYPE 2 ERRORS VS n

Hypothetical Impact of Tailored Phase II Trial Design on Patient Use in Phase III Studies

No. of Phase II Studies	Type I Error	Type II Error	No. of Positive Phase II Studies	No. of True-Positive Phase II Studies	No. of Patients per Phase III Study	Total No. of Patients in Phase III Studies
57.14	0.1	0.1	7.47	1.98	200	1,494
57.14	0.02	0.2	2.86	1.76	200	572
57.14	0.1	0.1	7.47	1.98	400	2,988
57.14	0.02	0.2	2.86	1.76	400	1,144
57.14	0.1	0.1	7.47	1.98	600	4,482
57.14	0.02	0.2	2.86	1.76	600	1,716
57.14	0.1	0.1	7.47	1.98	800	5,976
57.14	0.02	0.2	2.86	1.76	800	2,288

NOTE. Assuming that phase II studies are conducted using two-stage Simon optimal design with H_0 of 10% and H_1 of 30%, given a disease in which prior probability of success is 3.85%, a study using observed type I and type II error parameters will generate more false positives than true positives. In the scenarios presented here, a tailored phase II program accounting for the prior probability will reduce the No. of patients required for phase III studies by 62%. As the No. of patients required for a phase III study increases, the benefit of tailored trial design and reduction in false-positive phase II studies becomes larger, ranging from 922 if phase III studies have an average of 200 patients to 3,688 if phase III studies have an average of 800 patients.

By reducing the type 1 and type 2 errors from 0.1 and 0.1 to 0.02 and 0.2 respectively, the PPV of these studies would rise from 26.5% to 61.5%, , and the NPV would fall from 99.6% to 99.2%

WHAT ELSE MAKES POWERING DIFFICULT?

- Variation in the cohorts
- Bias e.g. selecting compounds with pre-specified ‘favourable’ phase 2 results and using these favourable results as the basis for treatment effect for phase 3 sample size planning*
- Measurements regarding observational occurrences or changes that are subject to bias:
  - Constitutional symptoms – fatigue, malaise, loss-of-appetite
  - Specific – pain, photophobia, parageusia
- Ordinal scales or ranking (evidence?)
- Variance in procedures e.g. timing and performance of collection, storage, testing clinical specimens
- Variance in assay performance (inter / intra-assay)
- Time (t) – at what threshold of time is a change relevant?*
- Delta – how much of a change is needed for relevance?

*Pharm Stat. 2012 Sep-Oct;11(5):373-85. doi: 10.1002/pst.1521. Epub 2012 May 29.

SIMPLE, EVERYDAY PROBLEMS WITH ‘n’

- Wrong null hypothesis
- Scaling – are we measuring on / off, yes / no or the significance of the data? i.e. is the effect of treatment big enough to make the intervention worthwhile, rather than does the treatment have any effect at all
- Most scientists think *p* tells them the probability the null hypothesis is true given their data…
- *p* actually tells us the probability of observing the data given that the null hypothesis is true. Something is ‘not guilty’ rather than innocent
- Noise – variance in the R (pre-study odds) will ultimately affect the required numbers to prove H_0 is true e.g. heterogeneity / homogeneity of populations studied e.g. serosusceptibility / Ab titres
- In a two sample situation, increasing the sample size of the intervention group to infinity does not send the power of the test to 1.0. The power will be limited by the sample size of the smaller group (e.g. placebo)
- The law of ‘diminishing returns’

SIMPLE, EVERYDAY SOLUTIONS TO 'n'

"…….it is estimated that 50% or more of all phase III trials performed are not successful. It can be argued that [phase II] assurance could provide a more realistic estimate of the probability of a trial's success."

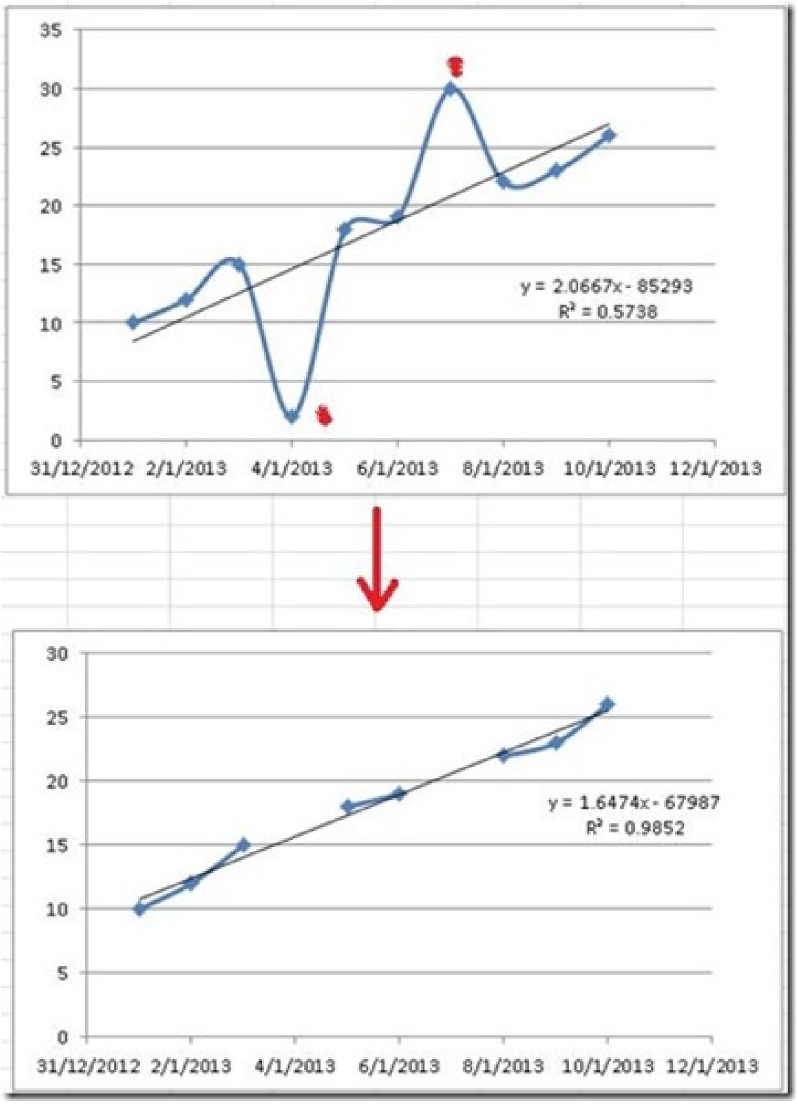
Kirby S, Burke J, Chuang-Stein C, Sin C. (Pfizer) Discounting phase 2 results when planning.

...no single solution to PhII predictability is the solution to PhIII performance (efficacy vs efficiency)

Powering considerations:

- Prospectively specify and rank all planned endpoints, time points, analysis populations and analyses
- Adjust cohort size so ‘n’ is large enough to take account of R, large/small confidence intervals, heterogeneity and variances in performance e.g. test specificity / sensitivity
- Factor in degree of control in a ‘controlled environment’ (vs type 1)
- PPV is highly sensitive to the variations in prior probability or odds (R)
- Phase III success rates seem to be related to P/N predictive values
- Type I and type II error rates in phase II are a major confounding factor in PhIII as they are amplified by n
- Reduce the multiplicity of endpoints – keep it simple – apply CI’s
- Composite scores (unweighted) may increase the likelihood of type 1 errors

BEWARE OUTLIERS / FRINGELIERS



- Outliers may be an indication of errors or unacknowledged variability
- Outliers can have deleterious effects on statistical analyses. First, they generally serve to increase error variance and reduce the power of statistical tests
- If something odd occurs 'more than seldom' it is a fringelier and may have significance

“In our laboratory (the Stanford Exploration Project or SEP) we noticed that after a few months or years, researchers were usually unable to reproduce their own work without considerable agony.”

- Claerbout describing his experiences in the mid-1980s

REDUCING THE NOISE SETTING A STANDARD

- Challenge trials (CHIMs) have simple, quantifiable and measurable 1^{o} endpoints
- CHIMs offer reduced noise by controlling the environment and reducing complexities of individuals, infection and disease
- R is known and characterized (FIH studies)
- Outliers are eliminated or reduced
- Powering calculations are simplified; cohort sizes are reduced in relation to increases in PPV

CHALLENGE VS TRADITIONAL PHII MODELS

Challenge study

- Small cohorts (50-100)
- Controlled environment
- High attack rate
- Known inoculation date
- Short duration (34-90d)
- Low cost (€2-3M)
- Early kill / no kill decisions
- May predict field trial design / performance
- Low noise / data ratio

Phase II field study

Large cohorts (250-300)
Uncontrolled environment
Low attack rate (prevalence)
Unknown inoculation date
Long duration (>1yr)
High cost (€5.5-6.5M)
Restricted window for enrolment
Extensive data analysis required for decisions
Large noise / data ratio

THANK YOU FOR YOUR ATTENTION

Life Science Services

Adrian Wildfire
Project Director
Infectious Diseases and HCU

SGS
BELGIUM NV
Generaal De Witterlaand, 19a, Bus 5
B-2800 Mechelen
BELGIUM

Mobile: +44 (0)7894 392625
Work: +44 (0)1483 828894
E-mail : adrian.wildfire@sgs.com
Web : www.sgs.com/lifescience

JOIN THE SCIENTIFIC COMMUNITY
CONNECT ON LINKEDIN
Discover and share current R&D market news and events including bio/analytical laboratory and clinical research drug development information.
www.sgs.com/LinkedIn-Life

CONTACT US

CLINICAL RESEARCH

lss.info@sgs.com

EUROPE : + 33 1 41 24 87 87
AMERICAS : + 1 877 677 2667

LABORATORY SERVICES

lss.info@sgs.com

EUROPE : + 41 22 739 9548
AMERICAS : + 1 866 SGS 5003
ASIA : + 65 637 90 111

www.sgs.com/lifescience

QUESTIONS ?