Randomised Controlled Trials in Mental Health- AFFIRM Short Course

Determining the Sample Size

Jenny Hellier

jennifer.hellier@kcl.ac.uk

Alan J. Flisher Centre for Public Mental Health

AFFIRM
AFrica Focus on Intervention Research for Mental Health

Determining the Sample Size

- Objectives
  - Examine why large trials are important
  - Assumptions and parameters of sample size calculations
  - Various methods available to calculate sample sizes in practice
  - Quick notes on errors
  - Practical – your turn!

Why do we need large trials

- Sample size matters – large RCTs are needed because , in general, we are looking to identify effects of an intervention that are often moderate in size (e.g. a 15 to 20% reductions in a substantive clinical outcome), usually a modest improvement over and above standard care

- Therefore there is clinical uncertainty about whether an intervention is effective

- Why do we need LARGE phase III trials?

- The smaller the treatment effect we want to detect, the more subjects we need

- Results from large trials statistically more reliable than from small studies
  - less likely to be due to chance (less random error)
  - results more precise (smaller confidence intervals)

- Phase II trials smaller – different methods for calculating sample size

Sample size

- Should be predetermined
  - Specified in grant application
  - Specified in the protocol
  - Specified in the main trial publication

  - Generally relates to a single primary outcome

  - Sample size is an educated guess!

Principle of sample size calculation

- The aim should be to have a large enough sample size to have a high probability (power) of detecting a clinically worthwhile treatment IF IT ACTUALLY EXISTS
- Larger studies have greater power to detect beneficial (or detrimental) effects
- Many clinical studies are far too small

Clinical Outcomes

- The trial should be adequately powered on the primary outcome

"Clinical outcomes should be expressed in a way that is most relevant to patients, even if such estimation reduces the statistical power of the study" [Rothwell. Lancet 2005; 365: 82-93]

- Sample sizes can be dramatically affected by the way in which the outcome is expressed

What information do we need

- Binary outcome
  - The effect size of interest
  - Anticipated proportion in each group
  - Values for the chosen significance (α)
  - And the chosen power (1-β)

- Continuous outcome
  - The effect size of interest
  - Anticipated mean in each group
  - The anticipated standard deviation (assumed to be the same in each group)
  - Values for the chosen significance (α)
  - And the chosen power (1-β)

Effect size of interest

- How do we determine the size of effect we wish to detect?

- 1) minimum clinically important difference
  - If the intervention groups differ by this much we can say it is likely to change clinical practice
  - Quite hard sometimes to establish

- 2) Data driven approaches
  - What is realistic given how much we know already?
  - Pilot study: standardised effect size , potential to influence a meta analysis

Smaller effect size → larger sample size

Effect size of interest

- Remember:

- Try and be as “realistic” as possible in the estimate of effect, do not over estimate the event rate or set up an implausibly large effect size

- Clinical Interpretation
  - Choice a scale that is clinically relevant
  - Looking for participant –centred outcomes

Significance Levels

- Also called alpha (α) level
- Usually set at 5% (0.05), but can be 1% or even 10%
- Definition: α level = type I error = you conclude there is a treatment effect when really, there is not
- So, you want this chance of a false positive result to be as small as possible
- **Smaller significance level → larger sample size**

The wrong conclusion 1

- In the context of a clinical trial concluding that there is a difference in outcomes between a treatment and control group, when in fact there isn't.

- As on the previous slide. A type I error
- Or (more usefully) a false positive

- Probability of making such an error is termed α, known as the significance level.

Power

- Power = probability you will conclude there is no treatment effect in your trial when there really is (a true positive)

- Power = 1-β = 1 - type II error

- Type II error = you conclude there is not a treatment effect when really there is (false negative)

- You want power to be high, so trials usually have a power of between 80% and 90%

Higher power → larger sample size

The wrong conclusion 2

- Falsely concluding that there is no evidence of a difference in the treatment and control group , when there is such a difference.

- Called a type II error
- Or (more usefully) a false negative

- Probability of making such an error is designated β, and 1 – β is commonly know as the statistical power

Scientific null hypothesis: no difference in response between two treatment groups

		Truth	
		Treatments are the same	Treatments are different
Clinical trial results	Treatments are the same	OK	Say a good treatment is no better than standard (type II error, β)
	Treatments are different	Say a worthless treatment is good (type I error, α)	OK (Power)

Choice of type I and II errors

- The choices of significance (α) and power (1-β) produce difference sample sizes:
  - Conventional: α = 5% and 1-β =80%
  - Optimal: α = 5% and 1-β = 90%
  - The best ! : α= 1 % and 1-β = 90%

- Consider the type of intervention, trial population available and the impact of a potential false-positive or false-negative finding

Continuous data

- Standardised effect size (Δ) = difference in means (δ)/ standard deviation (SD)

- n (per group) = 2 x k / Δ^2

- Where k = $[\ z_{(1-\alpha/2)} + z_{(1-\beta)}\ ]^2$

Sample size based on table of values

K	B= 0.10 Power 90%	B= 0.20 Power 80%
α = 0.01 (1%)	14.9	11.7
α = 0.05 (5%)	10.5	7.9

Note α is 2 sided

A far simpler way

- For $\alpha = 0.05$ and power of 80%
- $N \approx 31 / \Delta^2$ (total for 2 groups)

- For $\alpha = 0.05$ and power of 90%
- $N \approx 42 / \Delta^2$ (total for 2 groups)

- For $\alpha = 0.01$ and power of 90%
- $N \approx 60 / \Delta^2$ (total for 2 groups)

Continuous data an example

- Two arm RCT
- Outcome is depression measured by the Hospital Anxiety and Depression Scale (HADS)
- Mean in the placebo arm is expected to 15 points on the anxiety scale
- SD = 10 points
- Minimum clinical difference is deemed 5 points

- What is the sample size for $\alpha = 0.05$ and power of 80%

- $\Delta = \delta/SD = 5/10 = 0.5$

- Using the simple formula: $N(\text{total}) \approx 31/(0.5)^2 = 124$

Binary data

- Effect Size for for binary data
- Δ = $p_1 - p_2$ / √(ap (1- ap))
- Where ap is average proportion
- ap= $(p_1 + p_2)$ / 2

Binary data example

- Two arm RCT
- Outcome is abstinence from illicit drug use
- 20% of people are abstinence in the target trial population
- Reduction by 5% (20% - 15 %)

- What is the sample size for $\alpha = 0.05$ and power of 90%

- $\Delta = p_1 – p_2 / \surd(ap\ (1\text{-}\ ap))$

- $\Delta = 0.2 – 0.15 / \surd(\ 0.175(1\text{-}0.175)) = 0.1316$

- Using the simple formula: $N(total) \approx 42/\ (0.13)^2 = 2426$

General Guidelines

- In general, ≤ 0.20 is a small effect size,
- 0.50 is a moderate effect size
- and ≥ 0.80 is a large effect size (Cohen, 1992)

- d- standardized / Percentage of mean difference variance explained
  - Small 0.2 / 0 1%
  - Moderate 0.50 / 10%
  - Large 0.80 / 25%

Attrition

- Allowance for loss to follow-up, inflate the sample size
- Reasons: participants who don't comply with treatment, DNA scheduled visits / clinics , are lost to follow-up
- Typical not to get outcomes on all participants
- Inflate you sample size by 1/ (1-attrition proportion)
- For example
  - If you anticipate 15% loss to flow up (i.e. 0.15)
  - N * (1/(1-0.15)
  - N/ 0.85
- Will inflate your estimate accordingly

Clustering

- Between-cluster variation by $\sigma^2_{Between}$
- Within-cluster variation by σ^2_{Within}
- Intra-class correlation is given by $\rho_{Intra} = \frac{\sigma^2_{Between}}{\sigma^2_{Between} + \sigma^2_{Within}}$
- The sample size we calculate $m_{Individual}$ this assumes independent observations within each intervention group (ICC=0)
- To accommodate the clustering, $m_{Individual}$ needs to be inflated by our design effect (DE)
- K= number of people per cluster (c)

$$DE = 1 + (k - 1)\ x\ \rho_{Intra}$$

Randomistion Ratio

- May randomise more than 2 groups, adjust the sample size accordingly.
- May wish to randomise in different allocation ratios

"A total of 65 trials were identified; 56 were two-armed trials and nine trials had more than two arms. Of the two-arm trials, 50 trials recruited patients in favour of the experimental group. Various reasons for the use of unequal randomisation were given. Six studies stated that they used unequal randomisation to reduce the cost of the trial, with one screening trial limited by the availability of the intervention. Other reasons for using unequal allocation were: avoiding loss of power from drop-out or cross-over, ethics and the gaining of additional information on the treatment. Thirty seven trials papers (57%) did not state why they had used unequal randomisation and only 14 trials (22%) appeared to have taken the unequal randomisation into account in their sample size calculation"

1: Dumville JC, Hahn S, Miles JN, Torgerson DJ. The use of unequal randomisation ratios in clinical trials: a review. Contemp Clin Trials. 2006 Feb;27(1):1-12. Epub 2005 Oct 19. Review..

Randomisation Ratio

- If n is sample size required as per formula and n1 is the sample size for test and n2 is sample size for placebo and n1 /n2 = k, then
- n1 = (0.5) × n × (1 + k)
- n2 = (0.5) × n × (1 + (1/k))

- Suppose n = 100 and ratio between test (n1) and placebo (n2) is 2:1 (k=2) then
- n1 = (0.5) × (100) × (1+2) = 150
- n2 = (0.5) × (100) × (1+[1/2]) = 75

- As a statistician one should have clear and clinically meaningful inputs on the above points prior to working on sample size estimation.
- Generally, unbalanced designs require more subjects than the corresponding balanced designs.

Sample Sizes are a compromise

- Accruing several thousand patients in a ‘reasonable’ time may well pose a problem, even with collaboration through multi-centre studies

- Smaller trials are not worthless, but are unlikely to produce definitive results which will influence clinical practice

- In practice a balance is often imposed between the difference in response expected and the time / money / resources available for recruitment and trial management

Final word on sample sizes

- The greater the sample size → the greater the chance of showing a difference (if one exists)

- Large trials can detect smaller differences

BUT!!

- Differences must be clinically significant

Statistical significance is not the same as clinical significance!!

Software

- Formula
  - N query
  - G power
- Simulation
  - Stata
  - SAS
  - R
- Many online calculators / software (check the validation)

ARC sample size

- From meta-analysis of the core psychological interventions included in the PBI, it is estimated that the percentage of participants who are verified heroin and cocaine abstinent for the past 28 days (by self-report via structured clinical interview and individual diary record of heroin and cocaine use, confirmed by urine drug screen) at 22 weeks (14 weeks post-randomization) will be approximately 24% in the TAU and 42% in the PBI treatment arm (a relative risk of 1.75) (NICE 2008). This equates to an 18% difference in response status between the two trial arms (TAU and PBI) at week 22 (14 weeks post randomisation).

- The trial has been powered for an intention-to-treat analysis to evaluate the primary objective of the effectiveness of PBI verses TAU at 22 weeks (14 weeks post randomisation) respectively. There are 2 trial arms in the primary analyses: PBI intervention and TAU.

- Allowing for 16% attrition (This is the rate of attrition from opioid agonist treatments delivered across England obtained from the National Drug Treatment Monitoring System), an overall sample size of 368 participants, would be randomised to receive PBI or TAU, thus considering 184 undertaking BPI and 184 undertaking TAU, we would have 90% power to detect, the primary comparison, of a 18% difference in responder status at week 22 (14 weeks post randomisation) using 2 sided 5% significance tests. This calculation is based on the assumption of the proportion of participants who are verified heroin and cocaine abstinent for the past 28 days (by negative UDS) at 22 weeks (14 weeks post-randomization) will be approximately 24% in the TAU trial arm.

A final note on Errors

Two types of error:

- SYSTEMATIC ERRORS (i.e. biases)

and

- RANDOM ERRORS (i.e. chance)

We want these to be small, relative to the size of treatment effect that we expect to see.

Systematic Errors

- The only systematic difference between trial groups should be randomised treatment
  - Analysis can then find the difference between groups and assume this is due to randomised treatment
  - Any other systematic differences may mean the ‘treatment’ effect cannot be separated out from effect of other differences

- Steps to minimise bias:
  - Allocation concealment
  - Blinding (patients / doctors) to treatment allocation
  - Analysis by intention-to-treat

Random Errors

- Trials include samples of patients, not entire populations
- Results are estimates and not exact
- Measure random error in results using confidence intervals, p-values
- Sufficient sample size ➔ less random error in results

References

- Altman DG (1991). Practical Statistics for Medical Research. Chapman and Hall: London. Chapter 15; section 15.3 – sample size (p455 – 460)
- Kerry SM, Bland JM,. Statistical notes. Sample size in cluster randomisation. *BMJ* 1988; 316: 549
- Wittes J. sample size calculations for randomized trials. *Epidemiol Rev*. 2022; 24: 39-53