

# Designing powerful studies

Dominique Costagliola - Caroline Sabin

Why is it important to power a study correctly?

# Example 1

---

Two drugs (A and B) are compared in a randomised trial. The response rates in each group are:

(a)

Drug A            5/10

Drug B            6/10

Assuming all other factors are similar (e.g. side effects etc.) do you believe that drug B is more effective than drug A?

# Example 1

---

Two drugs (A and B) are compared in a randomised trial. The response rates in each group are:

	(a)	(b)
Drug A	5/10	50/100
Drug B	6/10	60/100

Assuming all other factors are similar (e.g. side effects etc.) do you believe that drug B is more effective than drug A?

# Example 1

---

Two drugs (A and B) are compared in a randomised trial. The response rates in each group are:

	(a)	(b)	(c)
Drug A	5/10	50/100	500/1000
Drug B	6/10	60/100	600/1000
	$p=0.653$	0.155	$<0.001$

Assuming all other factors are similar (e.g. side effects etc.) do you believe that drug B is more effective than drug A?

# Choice of sample size

---

- Thus when choosing the sample size for our studies we make a compromise between
  - Sufficient numbers to detect a treatment effect if it exists, and
  - Small enough numbers so that we don't waste resources or place too many patients at potential risk

# Errors in hypothesis testing

# Errors in hypothesis testing

	After carrying out RCT	
	$P < 0.05$ Conclude that new regimen is different	$P > 0.05$ Conclude no difference between regimens
New regimen really is different to existing regimen	✓	✗ Type II error ( $\beta$ )
New regimen is no different to existing regimen	✗ Type I error ( $\alpha$ )	✓



# Type I errors

---

- Every time a statistical test is performed, there is a risk that a Type I error will be made
- The  $P$ -value is the probability of obtaining the results by chance – this is the Type I error (a FALSE POSITIVE signal)
- All we can do to control the Type I error rate is to require stronger evidence (ie. a smaller  $P$ -value) before concluding significance
- We must be aware of Type I errors when interpreting the results of any study

# Example – 20 repetitions of a trial, no difference in outcome between regimens A and B

Trial no.	Regimen		P-value
	A	B	
	N <50 copies/ml	N <50 copies/ml	
1	28/54	22/46	0.84
2	24/53	26/47	0.42
3	30/61	20/39	1.00
4	25/51	25/49	1.00
5	29/57	21/43	1.00
6	24/50	26/50	0.84
7	22/51	28/49	0.23
8	30/54	20/46	0.32
9	28/57	22/43	1.00
10	20/47	30/53	0.23

Trial no.	Regimen		P-value
	A	B	
	N <50 copies/ml	N <50 copies/ml	
11	29/59	21/41	1.00
12	20/47	30/53	0.23
13	23/51	27/49	0.42
14	22/40	28/60	0.54
15	16/45	34/55	0.02
16	26/54	24/46	0.84
17	24/49	26/51	1.00
18	28/53	22/47	0.69
19	25/42	25/58	0.16
20	22/47	28/53	0.69

# Type II errors

---

- A Type II error occurs if you fail to reject the null hypothesis even if there is a true difference (a **FALSE NEGATIVE** signal)
- The major determinant of the Type II error rate is the size of the study
- Smaller studies are more likely to fail to detect a real effect than larger studies – increasing the size of the study will reduce the Type II error rate
- Variability is also a determinant when outcome is numerical

What is power?

# Power

	After carrying out RCT	
	$P < 0.05$ Conclude that new regimen is different	$P > 0.05$ Conclude no difference between regimens
New regimen really is different to existing regimen	✓ <b>POWER</b> (1- $\beta$ )	✗ Type II error ( $\beta$ )
New regimen is no different to existing regimen	✗ Type I error ( $\alpha$ )	✓

# Power

---

- The power of a study is the probability of correctly detecting a difference of any given size
- The power of a trial must be stated at the time of designing a trial as it will determine how many patients should be recruited
- The power will be low if the sample size is small - taking a larger sample will improve the power
- Ideally we would like a power of 100% but this is not feasible unless we recruit the entire population – we usually accept a power of 80-95%

# Interpreting power statements

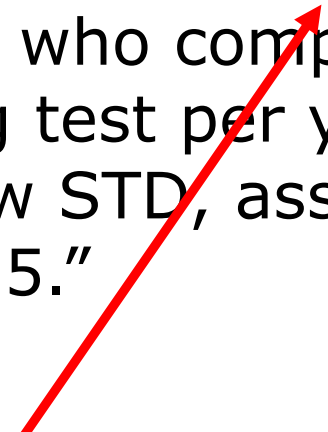
---

“The sample size was chosen as 400 in order to have sufficient power (0.8) to determine at least a 30% increase in the proportion of women who completed at least one asymptomatic screening test per year or in the proportion who acquired a new STD, assuming a 15% loss to follow-up and a of 0.05.”

# Interpreting power statements

---

“The sample size was chosen as 400 in order to have sufficient power (0.8) to determine at least a 30% increase in the proportion of women who completed at least one asymptomatic screening test per year or in the proportion who acquired a new STD, assuming a 15% loss to follow-up and a of 0.05.”




**If the intervention REALLY does lead to a reduction in the proportion of women with each outcome of 30% or more, there is an 80% chance that this would be detected as significant at the 5% level.**



# Interpreting power statements

---

“The sample size was chosen as 400 in order to have sufficient power (0.8) to determine at least a 30% increase in the proportion of women who completed at least one asymptomatic screening test per year or in the proportion who acquired a new STD, assuming a 15% loss to follow-up and a of 0.05.”




If the intervention REALLY does lead to a reduction in the proportion of women with each outcome of 30% or more, there is an **80% chance that this would be detected as significant at the 5% level.**

# Interpreting power statements

---

“The sample size was chosen as 400 in order to have sufficient power (0.8) to determine at least a 30% increase in the proportion of women who completed at least one asymptomatic screening test per year or in the proportion who acquired a new STD, assuming a 15% loss to follow-up and a of 0.05.”



If the intervention REALLY does lead to a reduction in the proportion of women with each outcome of 30% or more, there is an 80% chance that this would be detected as significant at the 5% level.

# Determining the size of a trial

# Determining the size of a trial (binary endpoint)

---

You need to know:

- 1) How many patients would expect to have the endpoint in the 'control' arm
- 2) What is the minimum 'treatment effect' that you would like to detect (ie. the smallest improvement in this proportion that is clinically important)
- 3) The type I error (usually 5%)
- 4) The power of the study (usually 80-95%)

# Example

---

- Randomised controlled trial of two drugs, A and B.
- Primary endpoint: proportion of patients experiencing virological response at 48 weeks
- 70% of patients receiving drug A (the standard of care) are expected to respond to therapy
- Would like to have 80% power to detect an improvement in response of 10% (ie. anticipated response rate in group B of 90%) at the 5% level of significance (type I error)

# Example

---

Significance level – 5%

Sample size required in each group if success rate on A = 70%

<b>Success rate on B</b>	<b>Number in each group when power equals:</b>		
	<b>80%</b>	<b>90%</b>	<b>95%</b>
<b>75%</b>			
<b>80%</b>		<b>392</b>	
<b>85%</b>			
<b>90%</b>			

# Example

---

Significance level – 5%

Sample size required in each group if success rate on A = 70%

<b>Success rate on B</b>	<b>Number in each group when power equals:</b>		
	<b>80%</b>	<b>90%</b>	<b>95%</b>
<b>75%</b>		<b>1674</b>	
<b>80%</b>		<b>392</b>	
<b>85%</b>		<b>161</b>	
<b>90%</b>		<b>82</b>	

# Example

---

Significance level – 5%

Sample size required in each group if success rate on A = 70%

<b>Success rate on B</b>	<b>Number in each group when power equals:</b>		
	<b>80%</b>	<b>90%</b>	<b>95%</b>
<b>75%</b>	<b>1251</b>	<b>1674</b>	<b>2070</b>
<b>80%</b>	<b>294</b>	<b>392</b>	<b>485</b>
<b>85%</b>	<b>121</b>	<b>161</b>	<b>199</b>
<b>90%</b>	<b>62</b>	<b>82</b>	<b>101</b>



# Determining the size of a trial (continuous, Normally distributed endpoint)

---

You need to know:

- 1) The mean value expected in the control arm
- 2) The minimum 'treatment effect' of interest (i.e. the smallest additional change in the measurement that is clinically important)
- 3) An idea of the variability associated with the measurement (e.g. standard deviation, variance)
- 4) The maximum type I error (usually 5%)
- 5) The power of the study (usually 80-95%)

# Notes on power calculations

---

- Power calculations are only a guide and are based on probabilities and assumptions
- Even in a well-powered study, an important effect may still be non-significant by chance
- Studies are usually powered for a single comparison between groups – multiple comparisons, sub-group analyses and multiple regression are likely to be underpowered
- Where the outcome is a rate, sample size calculations should also incorporate the length of the trial

# Allowing for loss to follow-up

---

- Even in ideal situations, it is unlikely that all patients will be followed without any loss to follow-up
- Can use methods to reduce the impact of loss to follow-up (e.g. ITT M=F)
- However, also sensible to increase sample size to allow for any loss to follow-up
- E.g. if 10% patients are expected to drop out of the study, then increase sample size accordingly – the number *after* drop-out should be the number obtained from the power calculation

# Example of a power calculation

---

“A total of 900 patients was needed to detect a difference at 3 years of 13% (from 50% to 63%) in the proportion of participants with HIV RNA below the limit of detection, or of 40 cells in the mean change in CD4 cell count between any two of the three groups with 80% power (5% significance) based on a global test on two degrees of freedom and assuming 10% loss to follow-up.”

## Example of a power calculation (2)

---

“We estimated that a sample of 6639 women would be needed to record 194 seroconversions, which was the minimum necessary for 80% power to detect a 33% reduction in risk of HIV seroconversion in Carraguard users, compared with placebo users. This calculation was based on a two-sided significance level of 0.05 and the following assumptions: (1) recruitment would last 18 months; (2) participants would be followed up for a maximum of 24 months or until a positive pregnancy or HIV test result, or both, or for a minimum of 12 months after enrolment of the last participant; (3) HIV incidence in the placebo group would be 3.5 per 100 woman-years; and (4) loss to follow-up would result in missing HIV outcomes for no more than 20% of participants.”

# ***Post-hoc* power calculations**

---

- Power calculations are based on assumptions
- These assumptions may be very different to the real-life situation – once a trial is finished we may realise that the original sample size was no longer sufficient
- If the treatment effect is not significant, it is tempting to perform a *post-hoc* power calculation
- However, the confidence interval for the treatment effect should tell us all we need to know about the power of the trial – a wide confidence interval suggests that the trial was insufficiently powered

# Reporting sample size calculations

---

- Review of 215 RCTs published in 6 high- impact general medical journals in 2005/2006
- 5% didn't report any sample size calculation; 43% didn't report all required parameters
- Difference between reported and replicated sample size was  $>10\%$  in 30%, and  $>50\%$  in 17% of trials in which sufficient data were available to re-calculate the sample size
- Only 34% trials reported all data required, had an accurate calculation, and used accurate assumptions for the control group

# Clinical versus statistical significance

---

- With a very large study even the smallest effect will be significant as the power will be high
- This does not mean that the effect will necessarily be of clinical relevance
- When assessing the results of any study, important to consider both the statistical (is the  $P$ -value  $< 0.05$  or does the confidence interval exclude 1 for a ratio or 0 for a difference?) and clinical (are the results important?) significance of the findings
- In a well powered study, findings that are clinically but not statistically significant should not occur



# Summary

---

- It is important to think carefully about power and sample size when planning a trial
- An awareness of the possible errors that can be made when carrying out a hypothesis test will be helpful when interpreting the results of the test
- Where possible, steps to reduce the chance of a Type I error (e.g. avoidance of unplanned interim or subgroup analyses) should be taken to ensure the reliability of the results