Confidence intervals

Caroline Sabin

Professor of Medical Statistics and Epidemiology
Institute for Global Health

Conflict of interest

I have received funding for the membership of Data Safety and Monitoring Boards, Advisory Boards and for the preparation of educational materials from:

- Gilead Sciences
- ViiV Healthcare
- Janssen-Cilag

Background

- Although *P*-values are helpful in telling us which effects are likely to be real, and which are likely to be chance findings, they suffer from several limitations
- In particular, the *P*-value by itself does not provide any helpful information about either the size of an association, or the impact of variability on this
- It does not allow us to put any findings into clinical context

Outline of session

- Some limitations of *P*-values
- How can confidence intervals help?

Outline of session

- Some limitations of *P*-values
- How can confidence intervals help?

Type I errors

- A *P*-value of 0.05 implies that there is a 5% probability that the results were due to chance
- For every 20 statistical tests we perform, we would expect that one of these would be falsely significant just by chance
- In this case, we would conclude that there was a real effect even though no effect exists
- This is a Type I error (a false positive finding)

100 trial participants - % women

Trial no.	Regimen		*P*-value
	A	B	
	N	N	
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	N	
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

Multiple testing

- Probability that ≥1 of our results will be falsely significant increases exponentially as the number of tests performed increases
- E.g. with 20 tests, the probability that *at least one of them* will have a *P*-value <0.05, even if there is no real effect, is almost 100%
- There are ways to deal with this (e.g. Bonferroni correction) but prevention is better than cure - focus on 1/2 key statistical tests, defined in advance and be wary of any presentation where a large number of *P*-values are presented

Example – dealing with multiple testing

- ACTG 5142 trial – comparison of three HAART regimens: EFV+2NRTIs; LPV/r+2NRTIs; LPVr+EFV+2NRTIs
- Three comparisons of interest
- Three planned interim analyses
- "The overall type I error rate was 0.05, with 0.017 (0.05 ÷ 3) allocated to each pairwise comparison between study groups; after adjustment for interim analyses, the final type I error rate was 0.014

Riddler SA et al. *NEJM* (2008); **358**: 2095-106

Limitations

- Small changes in the data can switch the results from being non-significant to significant

Limitations

	VL<50 copies/ml	VL >50 copies/ml	Total
A	11	25	36
B	45	42	87
Total	56	67	123

Chi-squared=3.79
P=0.0517

Limitations

	VL<50 copies/ml	VL >50 copies/ml	Total
A	11	25	36
B	45	42	87
Total	56	67	123

Chi-squared=3.79
P=0.0517

	VL<50 copies/ml	VL >50 copies/ml	Total
A	11	26	37
B	45	41	86
Total	56	67	123

Chi-squared=4.45
P=0.0348

Limitations

- Small changes in the data can switch the results from being non-significant to significant
- Threshold of 0.05 is rather arbitrary – what do you do if *P*=0.05? Is this significant or non-significant?

Limitations

- Small changes in the data can switch the results from being non-significant to significant
- Threshold of 0.05 is rather arbitrary – what do you do if *P*=0.05? Is this significant or non-significant?
- If study is large enough, results can be statistically significant even if not clinically important

Limitations

	VL<50 copies/ml	VL >50 copies/ml	Total
A	750 (75%)	250 (25%)	1000
B	770 (77%)	230 (23%)	1000
Total	1520	480	2000

Chi-squared=0.99
P=0.32

Limitations

	VL<50 copies/ml	VL >50 copies/ml	Total
A	750 (75%)	250 (25%)	1000
B	770 (77%)	230 (23%)	1000
Total	1520	480	2000

Chi-squared=0.99
P=0.32

	VL<50 copies/ml	VL >50 copies/ml	Total
A	7500 (75%)	2500 (25%)	10000
B	7700 (77%)	2300 (23%)	10000
Total	15200	4800	20000

Chi-squared=10.86
P=0.001

Outline of session

- Some limitations of *P*-values
- How can confidence intervals help?

Treatment effects

- *P*-values by themselves are of limited value
- Although they give an indication of whether the findings are likely to be genuine, they do not allow you to put findings into clinical context
- Should provide an estimate of the effect of interest (i.e. some comparative effect) as well as an indication of the precision of the estimate (i.e. its 95% confidence interval)

Treatment effects

- The 'treatment effect' is the additional benefit that the new drug/regimen provides compared to 'standard of care'
- Example:
  - Drug A (standard of care) 63% response
  - Drug B (new regimen) 71% response
- The treatment effect is 8% (= 71% - 63%)
- For every 100 patients treated with regimen B, expect that an extra 8 patients would respond, compared to the number that would have been expected had they been treated with regimen A

How do we interpret trial outcomes?

- Estimate of 8% was a point estimate; this is our 'best guess' but it gives no indication of variability
- Confidence intervals provide a range of additional plausible values that are supported by the results of the study – they indicate the precision of the estimate
- In a trial, the 95% CI for the treatment effect allows us to put the results from the trial into clinical context; can weigh up benefits in light of any disadvantages of drug (e.g. increased cost or worse toxicity profile)

Example

Trial number	Drug				
	A		B		
	n	n (%) responding	n	n (%) responding	Difference (B – A)
1	50	34 (68)	50	40 (80)	12%

- We believe that drug B is 12% more effective than Drug A
- The 95% CI for this estimate is: -5.0% to +29.0%
- Drug B could be up to 5% *less effective* than drug A, or up to 29% *more effective* than drug A
- What are your views about drug B?

Example

Trial number	Drug				
	A		B		
	n	n (%) responding	n	n (%) responding	Difference (B – A)
1	150	102 (68)	150	120 (80)	12%

- We believe that drug B is 12% more effective than Drug A
- The 95% CI for this estimate is: 2.2% to 21.8%
- Drug B could be as little as 2% *more effective* or as much as 22% *more effective* than drug A
- What are your views about drug B?

Precise vs imprecise estimates

- First confidence interval was too wide to allow us to judge whether drug B was better, worse or the same as drug A
- The estimate was imprecise, or lacked precision
- Second confidence interval was narrower, allowing us to conclude that drug B was likely to be better than drug A
- The estimate from this trial was more precise
- Major determinant of width of CI is the sample size

How do you obtain a narrower CI?

Assume that 68% of patients on drug A and 80% of patients on drug B respond to therapy….

Number in each group	Treatment ‘effect’	95% CI for treatment effect
50	12.0%	-5.0%, +29.0%

How do you obtain a narrower CI?

Assume that 68% of patients on drug A and 80% of patients on drug B respond to therapy….

Number in each group	Treatment 'effect'	95% CI for treatment effect
50	12.0%	-5.0%, +29.0%
100	12.0%	-0.0%, +24.0%
150	12.0%	+2.2%, +21.8%
200	12.0%	+3.5%, +20.1%
300	12.0%	+5.1%, +19.0%
500	12.0%	+6.6%, +17.4%

Other points

- Although we have focussed on confidence intervals for the difference in two proportions, they can be generated for almost every statistic
- Calculations may be tricky, but most statistical packages will generate them automatically
- Most journals now require that confidence intervals are provided for all treatment effects reported in a paper

Summary

- We use *P*-values to judge whether any effects we see are bigger than would be expected by chance
- However, they suffer from a number of limitations so should not be interpreted in isolation
- Any comparison should always be accompanied by some measure of effect size (e.g. the difference in proportions with a virological response) and a confidence interval for this effect
- For some types of RCT, such as equivalence or non-inferiority trials, confidence intervals are even *more* important than *P*-values

Over to you…