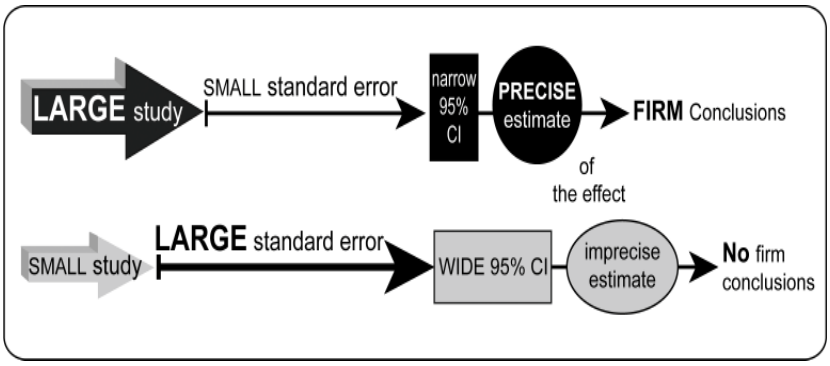
Small studies strengths & limitations

How small is “small”

- There is no rule over what defines a small study.
- N=20 could be sufficient in some situations (eg if looking for a very large difference or correlation), but far too small at other times.
- Regardless of the effect of interest, the larger the study the smaller the standard error; and the latter is what we ultimately want



But remember that a large standard error could also be found if there are few events (counting people/time-to-event endpoints), or large standard deviation (taking measurements)

Strengths

- Study can be quick to do
- Address the study question in a shorter space of time
- Only need a few centres (if relevant)
- Obtaining ethical and other approvals can sometimes be easier (just need local/single centre approval)
- You can test your hypothesis in a smaller sample first; if negative, you haven't spent too much resources on looking for an association that doesn't really exist (ie you save time, money, lab materials, patients etc)

Limitations

- Finding and interpreting p-values when not small
- Interpreting confidence intervals
- Finding spurious (unexpected, unusual or annoying) results. But you can explain this away by saying you have a small study (more difficult to say this with a large study!)
- If the spurious finding looks interesting, discuss it, but make clear that it came from an "exploratory analysis", and wasn't part of your original hypotheses/objectives
- Can sometimes get the wrong answer, or they over-estimate the effect

- Eg, suppose we want to know the smoking prevalence in a group of people
- If N=20 and 5 say they smoke, we estimate the prevalence to be 25% (5/20)
- However, we can see in this situation that N=20 is not a very reliable sample size

- If we just happen to have 2 fewer smokers, we'd estimate the prevalence to be 15%
- If we just happen to have 2 more smokers, we'd estimate the prevalence to be 35%
- Chance variability (natural variation, bad luck) can lead to quite different quantitative estimates

- 95% CI is 9 to 49%
- We think the true prevalence could be anywhere between 9% (a low number) and 49% (a fairly high number)
- Here, the 95% CI is not very helpful/informative (because of the small sample size)

Randomised blind trial of 2 interventions in treating mild/moderate depression

Average depression scores	Intervention A N=13	Intervention B N=16	Placebo N=16
Baseline	11.85	11.37	10.43
6 weeks later	4.69	3.06	8.5
% reduction	60%	73%	18%
P-value	P=0.05	P=0.004	

A standard and well-established questionnaire was used to assess depression

Did either intervention A or B work?

Would you recommend either intervention?

Randomised blind trial of Reiki in treating depression

Average depression scores	Hands-on Reiki N=13	Distance Reiki N=16	Placebo N=16
Baseline	11.85	11.37	10.43
6 weeks later	4.69	3.06	8.5
% reduction	60%	73%	18%
P-value	P=0.05	P=0.004	

Does Reiki work?

Randomised blind trial of Reiki in treating depression

Average depression scores	Hands-on Reiki N=13	Distance Reiki N=16	Placebo N=16
Baseline	11.85	11.37	10.43
6 weeks later	4.69	3.06	8.5
% reduction	60%	73%	18%
P-value	P=0.05	P=0.004	

Problems: small trial; no scientific basis for an effect
It could be a fluke result, even though 'statistically significant' (ie it could be one of the 4 in 1000)
All the trial subjects responded to an advert for the trial, so are probably more likely to show a placebo effect
Therefore, the apparent overall reduction in the mean scores, could partly be due to big (chance?) improvements in only 1 or 2 patients and little difference in the others – we need to look at scatter plot (ie did all/most patients show improvement or only 1-2?)

Many researchers do not fully understand what p-values really mean, though they are found throughout most journal articles
And interpreting research studies is often (incorrectly) focussed on the p-value

Writing up small studies

- Acknowledge if it is smaller than it should be
- If appropriate, do a sample size calculation and talk about it, but base the effect size on evidence published before your study. Eg:
- *"Smith et al 2002 suggested that the relative risk could be 0.75. The sample size needed to detect this is 250 patients, with 80% power and 5% significance level. In the time allowed, we could only recruit 100 patients, so we are aware that our study is underpowered. However, our results are consistent with……so despite not reaching formal statistical significance, the data are suggestive of an effect….."*

Writing up small studies

- Do not do a sample size calculation based on your observed result, even if you get a statistically significant result.
- Some supervisors/examiners think this is a good idea, especially if a p-value is not statistically significant but there seems to be a meaningful effect (ie it's their attempt to explain away the lack of a small p-value).
- But this approached is biassed and doesn't really have much meaning (the study has been done). If you didn't find a small p-value in the first place then you'd expect the study to be too small/underpowered!
- References:
- Hoenig & Heisey. The abuse of power: the pervasive fallacy of power calculations for data analysis. American Statistical Assoc 2001; 55(1): pp19-24
- Goodman & Berlin. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. Annals Internal Medicine 1994; 121: pp200-206

- Don't make overly strong conclusions when a result is unexpected, even if you find a highly statistically significant result (eg $p<0.001$)
- It could be too good to be true (eg lung cancer example above)
- Being honest about the study limitations is always better than making exaggerated claims (and is usually a sign of a more thoughtful researcher!)
- Always drive the interpretation around what you consider to be clinically important and whether the point estimate of the effect (eg relative risk, mean difference, hazard ratio, or correlation) is consistent with your prior hypotheses or other people's work.
- Then consider 95% CIs
- Then consider p-values
- If appropriate, be clear that your findings should be confirmed in other, larger studies

Finding unexpected results

- Reasons for an unusual/unexpected result:
  - It's real
  - It's spurious (you can't find an explanation)
  - It's spurious (but can be explained away)
- Can find unexpected results in small or large studies
- Indeed, large studies often have many variables, and researchers are too tempted to analyse the data in lots of ways (believing that the study size will produce reliable data).

Before further investigation of data

- Check your coding in the stats software!
- Make sure everything is labelled the correct way round. Eg, if comparing males (code=0) and females (code=1), some stats packages will automatically make code=0 the comparison group; others will use code=1
- Check finding with supervisor
- Look at literature:
  - same exposure, same population
  - Same exposure, different population
  - Similar exposure, same population

PSA screening for prostate cancer

	No. screened	No. deaths from prostate cancer	Relative risk (95% CI) of dying from prostate cancer in screened group vs control
USA	77,000	174	1.11 (0.83-1.50)
Europe	162,000	540	0.80 (0.67-0.95)

What do you conclude from these 2 trials?

Does PSA screening work or not?

Results from 2 trials

	No. screened	No. deaths from prostate cancer	Relative risk (95% CI) of dying from prostate cancer in screened group vs control
USA	77,000	174	1.11 (0.83-1.50)
Europe	162,000	540	0.80 (0.67-0.95)

- In the US trial, 15% of men in the screened group declined, and as many as 50% in the control group had PSA testing (this would dilute the effect of screening)
- Also, 44% of men already had PSA testing before the trial (so cancers found during the trial will tend to be those not easily found by PSA)
- Both trials show no mortality benefit in first 7 years, but possibly 50% reduction after 10 years

Interpreting unexpected findings

- Don't automatically think you're the first to find some wonderful result; you may have done, or it might be completely spurious
- Look for prior evidence on this, including speaking to supervisor/colleagues
- Do additional data analyses, but always be clear that if you do think you've found an explanation, it could still be a spurious finding in your one study
- Always consider plausibility
- Many things in medicine, which we now take for granted, started off as "spurious" but interesting findings, that led to further and confirmatory work.
- Don't use language that implies your conclusions are certain; try to provide some cautionary notes.

When studies are not quite big enough

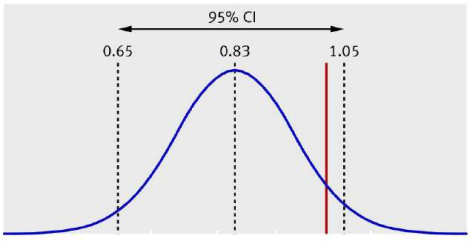
- The EICESS-92 phase III trial, based on high-risk patients with Ewings sarcoma, aimed to determine whether adding etoposide to standard chemotherapy would improve event-free survival (the chance of cancer recurrence or death).
- Powered to detect a hazard ratio (HR) of 0.60 (40% relative risk reduction), the target sample size was 400 patients (492 were recruited).

- But observed HR was **0.83, 95% CI 0.65 to 1.05, p=0.12**.
- What do you conclude?
- Because p>0.05 it would normally be concluded that there is insufficient evidence for an effect.
- However, the observed 17% risk reduction is clinically important, though smaller than expected, 40%.

- Most researchers understand that the true effect is likely to lie *somewhere* in the confidence interval, hence the possibility of it being one (ie no effect).
- However, there is a common misconception that the true effect lies anywhere within this range *with equal likelihood*.

95% Confidence Interval

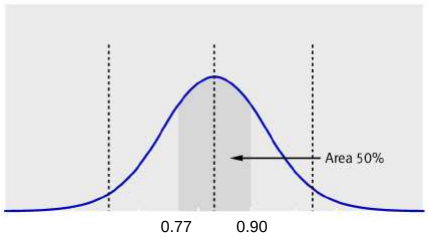
- The true HR is more likely to lie around the middle, i.e. estimated HR (0.83), than at the extremes of the confidence interval.



The 95% CI and the no effect value (hazard ratio = 1) Although the CI includes 1, most of the range is below it

95% Confidence Interval – Ewings sarcoma trial

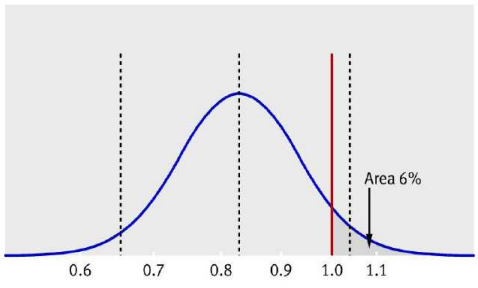
There is a 50% chance that the range 0.77 and 0.90 contains the true hazard ratio



Similarly there is a 75% chance that 0.72 and 0.95 contains the true HR.

95% Confidence Interval

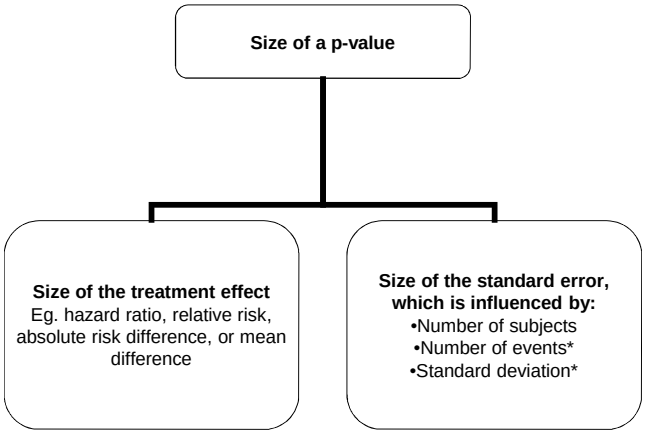
- The upper limit of the confidence interval is 1.05 and only just exceeds 1.0.



There is only a 6% chance that the range ≥1.0 contains the true HR

- The conclusion reported in the paper was that "the addition of etoposide *seemed* to be beneficial".
- This is the only randomised trial of etoposide in these children.
- The disorder is uncommon: 6.5 years to recruit 492 patients across Europe. Another trial is unlikely.
- Although the target sample size was exceeded, the treatment effect was smaller than expected (HR 0.83 vs 0.60), hence why the result was not statistically significant (i.e. trial was not big enough).

Remember



- Once a study has finished, there is no going back!
- The researchers are then stuck with whatever results arise
- Most see no problem with claiming a treatment effect, when, for example
- Relative risk is 0.75, 95% CI 0.57-0.99, and p=0.048
- But what about: 0.75, 95% CI 0.55-1.03 and p=0.07??
- These 'borderline' p-values are not uncommon

- In 6 major medical journals reviewed in 2009, 24 out of 287 phase III trials (1 in 12) had borderline results:
- 0.05<p-value<0.10 or a lower/upper 95% CI close to the no effect value (but just exceeding it)

Example 1

Interventions and patient group	Primary endpoint	Main result	Conclusion reported in the Abstract
Nurse-led psycho-educational intervention versus usual care for palliative care in patients with advanced cancer N=322	Symptom intensity (measured on a continuous scale)	Mean difference: -27.8 scores (95% CI -57.2 to +1.6) P=0.06	*"Those receiving nurse-led… intervention…… did not have improvements in symptom intensity scores"*

Bakitas et al, *JAMA 2009;302:741-9.*

Example 2

Interventions and patient group	Primary endpoint	Main result	Conclusion reported in the Abstract
Tailored care plan versus usual care in patients with coronary heart disease N=903	Patients with systolic blood pressure >140mm Hg at 18 months (hospital admission was another endpoint)	Odds ratio 0.66 95% CI 0.43 to 1.01 P=0.06	"*Admissions to hospital were significantly reduced…but no other clinical benefits were shown*"

Murphy et al, *BMJ 2009;339:b4220.*

Example 3

Interventions and patient group	Primary endpoint	Main result	Conclusion reported in the Abstract
Pre-surgical chemoradiotherapy versus chemotherapy in patients with locally advanced cancer of the esophagogastric junction. N=126	Overall survival	Hazard ratio 0.67 95% CI 0.41 to 1.07 P=0.07	"*Although… statistical significance was not achieved, results point to a survival advantage for preoperative chemoradiotherapy*"

Stahl et al, *J Clin Oncol 2009;27:851-6.*

Example 4

Interventions and patient group	Primary endpoint	Main result	Conclusion reported in the Abstract
Aerobic exercise training plus usual care versus usual care alone, in patients with chronic heart failure N=2331	All-cause mortality or hospitalisation	Hazard ratio 0.93 95% CI 0.84 to 1.02 P=0.13	*"…exercise training resulted in non-significant reductions in the primary endpoint…."*

O'Connor et al, *JAMA 2009;301:1439-50.*

Example 5

Interventions and patient group	Primary endpoint	Main result	Conclusion reported in the Abstract
Artesunate suppository versus placebo in patients with severe malaria who cannot be treated orally; N=12,068	Mortality	Risk difference -0.4% 95% CI -1.0 to +0.2% P=0.1	*"…..a single inexpensive artesunate suppository… substantially reduces the risk of death or permanent disability"*

Gomes et al, *Lancet 2009;373:557-66.*

Example 6

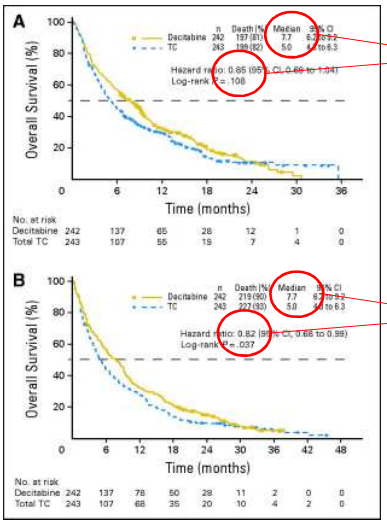
Interventions and patient group	Primary endpoint	Main result	Conclusion reported in the Abstract
Telephone counselling using cognitive behavioural skills vs. no intervention to encourage smoking cessation in adolescents; N=2151	6-months prolonged abstinence from smoking	Absolute risk difference 4.0% 95% CI -0.2 to 8.1% P=0.06	*"…personalized motivational interviewing...is effective in increasing teen smoking cessation"*

Peterson et al, *J Natl Cancer Inst 2009;101:1378-92.*

What language to use for borderline results

- Don't say with certainty that there is or isn't an effect
- Borderline p-values (>0.05 but <0.10), do not give strong evidence either for or against an effect
- Use words like 'suggestion of an effect'; also 'indication', 'seems', and 'trend' (though be aware that 'trend' is also used for other things)
- Discuss the strengths and limitations of the result, and try to back it up with other evidence if possible
- Perhaps conclude with a recommendation for further confirmatory studies
- See Hackshaw & Kirkwood, BMJ 2011 for more details

Overall survival: Anti-cancer drug, decitabine, for acute myeloid leukemia



Survival time increased by 2.7 months;
Risk of dying decreased by 15%
P-value = 0.108

Spot the difference

Survival time increased by 2.7 months;
Risk decreased by 18%
P-value = 0.037

Kantarjian H M et al. JCO 2012;30:2670-2677

Remember, same trial:

- FDA press release, Feb 2012
- *"The study **failed to show that patients on decitabine lived any longer** than patients in the control group, the FDA reviewers said.*

- EMEA press release, Oct 2012
- *"The European Commission has approved decitabine for acute myeloid leukemia (AML) in adults 65 years and older….. The protocol-specified final analysis demonstrated a 54% increase in median overall survival in patients in the decitabine group, compared with those in the comparator group"*

	No. of patients	No. of deaths	Reduction in risk of dying	Difference in median survival
Target	480	385	25%	2 months
Observed	485	396	15%	2.7 months
		446 (updated)	18%	2.7 months

They exceeded the target for the median survival.
But the problem is that the sample size is based on the target % reduction in risk of 25%
Main treatment effect (risk of dying) is smaller than expected (15% instead of 25%),
therefore a larger trial would have been required

If you cannot get a larger study, then an alternative is to get longer follow up.
Remember: number of events is more important than number of people

Therefore, in this study, the result based on the largest number of deaths should be more reliable

The above example is based on a clinical trial, but the principles are the same for any study type