

Clinical versus Statistical Significance

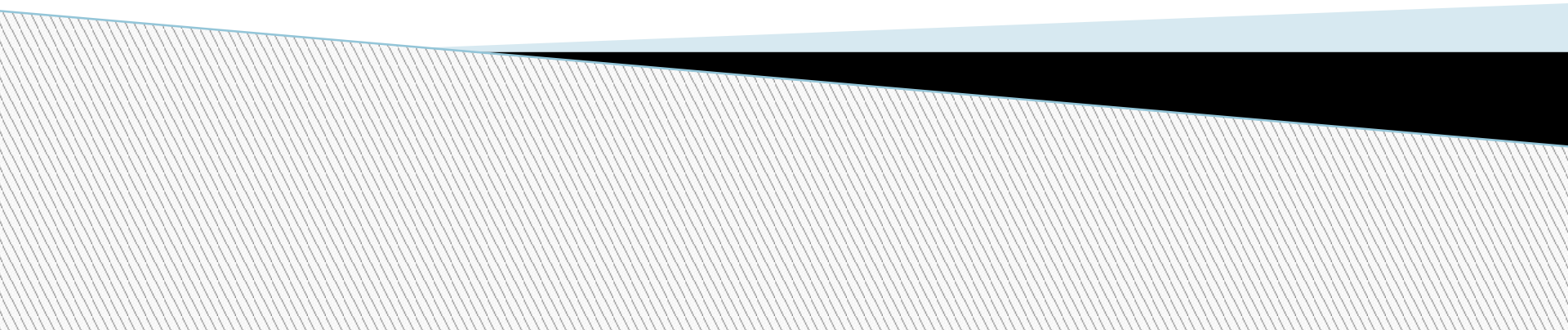
Jonathan Levin MSc, MSc, PhD

Division of Epidemiology and Biostatistics

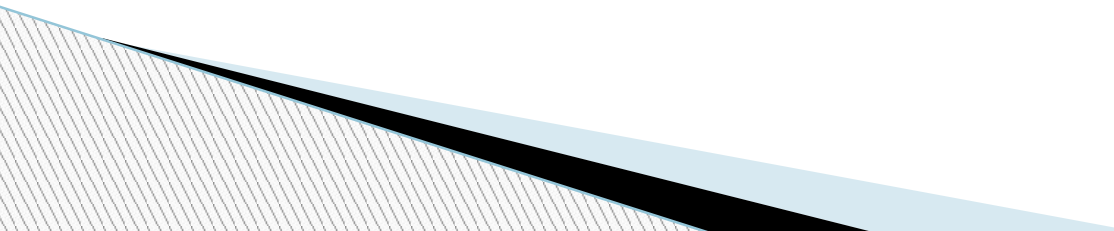
Wits School of Public Health

Prof Elena N Libhaber MSc, MSc, PhD

Health Science Research Office



Outline

1. Problem Statement
 2. Decision rule – BHSc Hons results
 3. Hypothesis testing & statistical significance
 4. Hypothesis testing & confidence intervals
 5. Example 1 – PCI in MI
 6. Example 2 – Thai HIV-1 vaccine trial RV144
 7. Conclusions
- 

Problem Statement

We are comparing an outcome between two groups (e.g. FEV1 between standard and new asthma inhaler)

Is there a real difference in mean FEV1 between the two groups ?

(Statistical significance)

Is this difference large enough to change clinical practice (e.g. switch to new inhaler)

(Clinical significance)



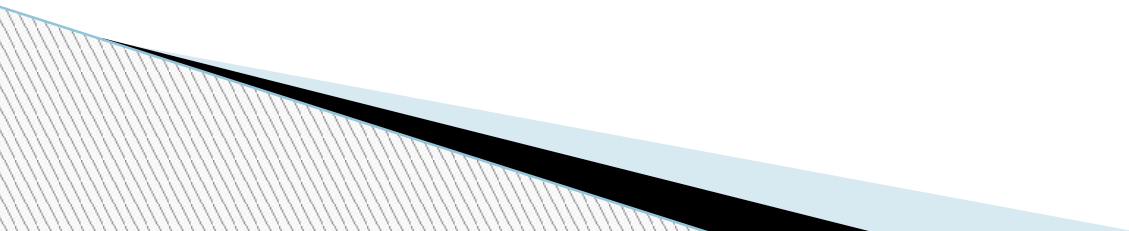
Decision Rule – BHSc (hons)

A student requires 50% (overall) in order to pass

A mark of 75% or above gains a distinction

A mark of between 40% and 49% earns a supp

A mark of below 40% is a fail and student must repeat



Decision Rule – BHSc Hons

Student gets 78% - pass with distinction

Student gets 68% - pass

Student gets 53% pass (??? Admit to MSc)

Student gets 47% supp

(Not that different to previous student)

Student gets 36% - fail


Hypothesis testing & statistical significance

A hypothesis test measures the evidence against a null hypothesis usually of no effect

(e.g. there is no difference between the mean FEV1 in the two groups)

Finding is summarised as a P-value

“Probability of getting a result as extreme or more extreme than that observed if null hypothesis is true”



Hypothesis testing & statistical significance

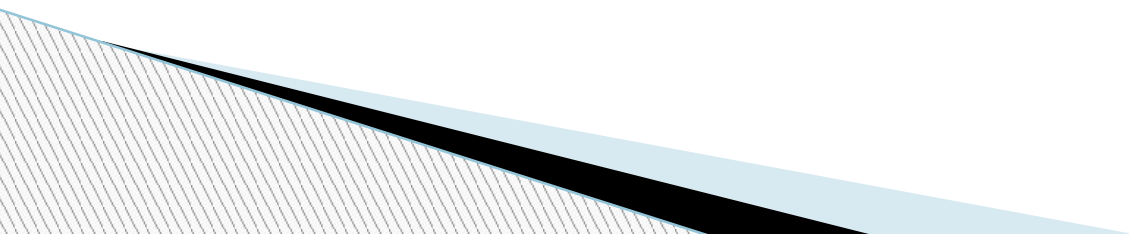
Understanding the p- value:

What is Hypothesis testing?

- 1) Null hypothesis Ho of a statistical test: hypothesis of no differences

p-value refers to Ho: rejecting or not rejecting Ho

- 2) The alternative hypothesis H_a or the hypothesis of the researcher is the opposite of the H_o
(which is what we expect or hope to be true.....)



Hypothesis testing & statistical significance

Decision rule based on P-value from hypothesis test

1. If $P > 0.10$ then there is no evidence of an effect
2. If $0.05 < P < 0.10$ (trend towards significance/
marginal significance) – worth another look
3. If P lies between 0.045 and 0.049 – statistically
significant but need to do further work
4. If $P < 0.045$ (even if $P < 0.001$) interpret magnitude of
difference – “effect”

Hypothesis testing & confidence intervals

One aid to interpreting the magnitude of the difference is a confidence interval

“ Imperfectly understood CIs are more useful and less dangerous than incorrectly understood p- values”

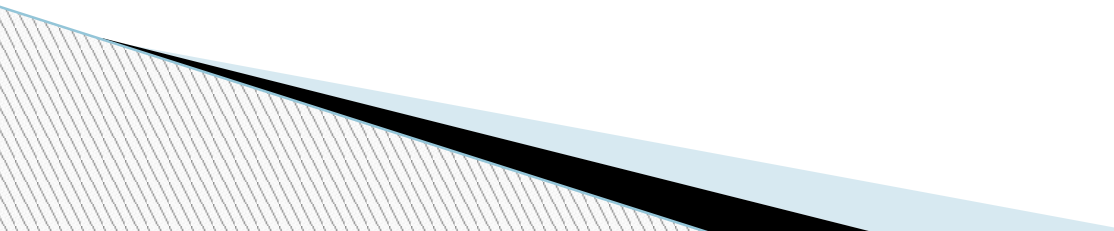
Hoening and Heisey (2001)



Hypothesis testing & confidence intervals

When comparing two means we can reject the null hypothesis of no difference between the means if the 95% confidence interval (CI) for the difference does not include 0 – so can use confidence intervals to carry out hypothesis tests

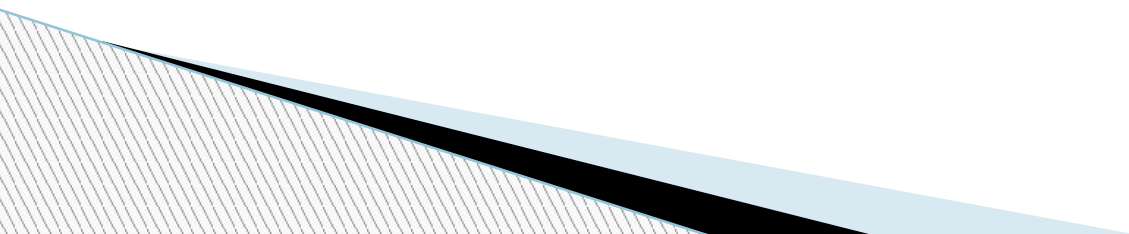
In addition the CI shows the magnitude of any difference



Hypothesis testing & confidence intervals

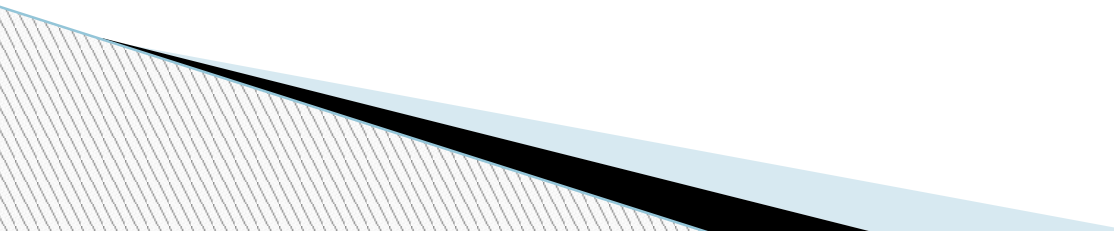
For comparing two means the confidence interval and hypothesis test will give identical results

In other cases the results are very similar (but not necessarily identical)



Hypothesis testing & confidence intervals

Ex: (Hypothetical) It is hoped that by the third trimester 90% of pregnant women will have had at least one antenatal visit. In an informal settlement it was found that 170 / 200 pregnant women had at least one antenatal visit before the third trimester. Is there evidence that this differs from the expected proportion?



Hypothesis testing & confidence intervals

$$n=200 \quad nP=200*0.90 = 180 \quad nQ=200*0.10= 20$$

So the normal approximation is valid

We test $H_0 : P = 0.90$

vs $H_1 : P \neq 0.90$

$$p = 170 / 200 = 0.85$$

Note that $P_0 = 0.90$ so $Q_0 = 1 - 0.90 = 0.10$ and the standard error of P (proportion who have had some antenatal care) is given by

$$\sqrt{\{PQ / n\}} = \sqrt{\{0.90*0.10 / 200\}} = 0.0212$$

Hypothesis testing & confidence intervals

Thus our test statistic is

$$Z = \{ p - P_0 \} / \text{se}(p) = \{0.85 - 0.90\} / 0.0212 = -2.36$$

Critical points of the standard normal distribution are
5% - 1.96 and 1% - 2.58

Thus we can reject H_0 at the 5% level: there is strong evidence that women from the district where our sample was drawn have a lower uptake of prenatal care

P-value from tables is $P=0.018$



Hypothesis testing & confidence intervals

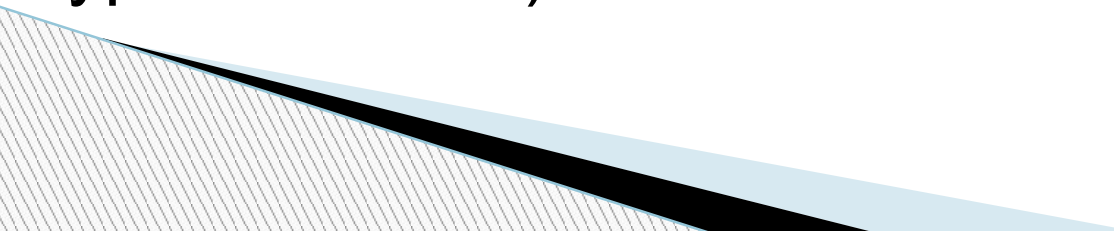
Can find 95% confidence interval for proportion of women in informal settlement who have had antenatal care by third trimester

In this case we do not know P so we estimate the standard error of p using

$$\text{s.e. } (P) = \sqrt{\{pq/n\}} = \sqrt{\{0.85 \times 0.15 / 200\}} = 0.0252$$

Then our 95% confidence limits for P are given by $p \pm 1.96 \text{ se } (p)$ or $(0.801 ; 0.899)$

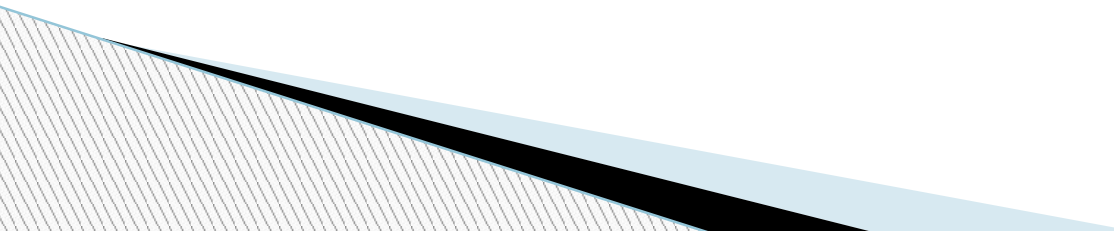
(Note that the s.e. is different from that of the hypothesis test)



Hypothesis testing & confidence intervals

We are “95% sure” that the true proportion of women in the informal settlement who have had some antenatal care by the third trimester is between 0.801 and 0.899 i.e. between 80.1% and 89.9%

It is less than 90% (at least 0.1% less) but the difference is not practically (clinically) significant



Hypothesis testing & confidence intervals

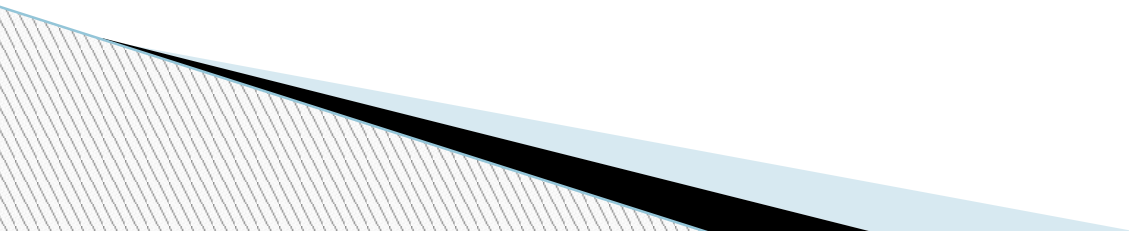
Confidence intervals are often used to formally carry out hypothesis tests e.g. in equivalence and non-inferiority trials

In the DART trial (Uganda & Zimbabwe) Clinically driven monitoring (CDM) would be judged to be non-inferior to routine lab monitoring (LDM) if the upper 95% confidence limit for the hazard ratio was less than 1.17 (i.e. it did not lead to more than a 17% increase in the risk of an endpoint)

Hypothesis testing & confidence intervals

How can we decide if a difference is clinically significant?

For continuous outcome measures Cohen's idea of effect size is important



Hypothesis testing & confidence intervals

Cohen effect size=

change in outcome variable experimental group vs control

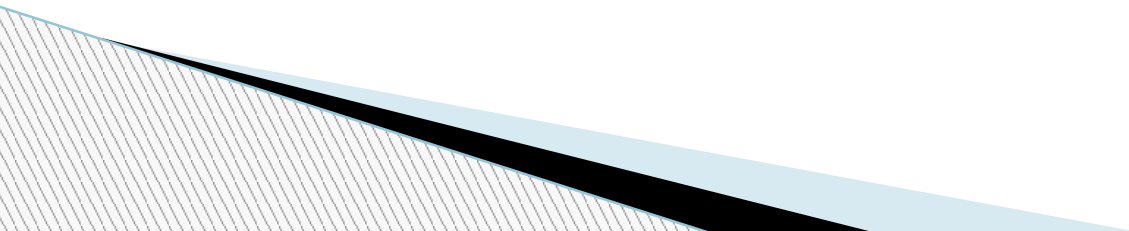
Standard deviation of both groups (pooled SD)

<0.2 trivial effect

0.2-0.5 small effect

0.5-0.8 moderate effect

>0.8 large effect

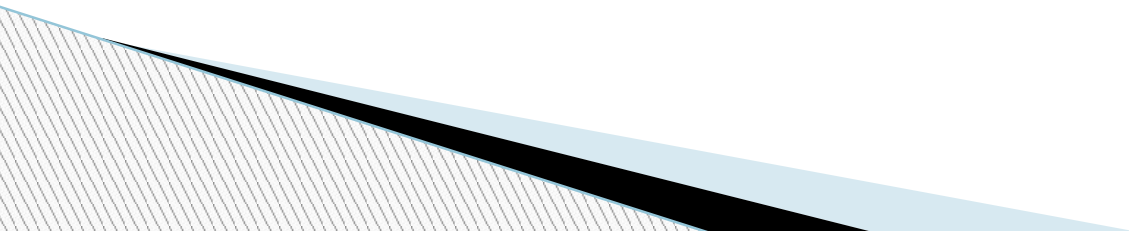


Hypothesis testing & confidence intervals

For ratio measures (e.g. OR or RR) follow advice of the late Prof Syd Shapiro

“Epidemiologist counts 1, 2, big”

So OR or RR > 2 is important

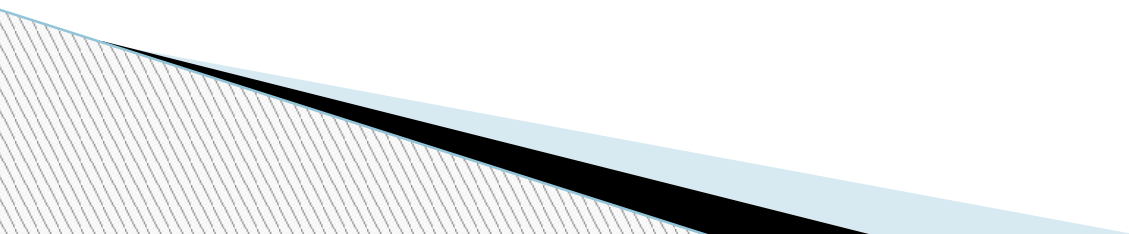


Hypothesis testing & confidence intervals

NB

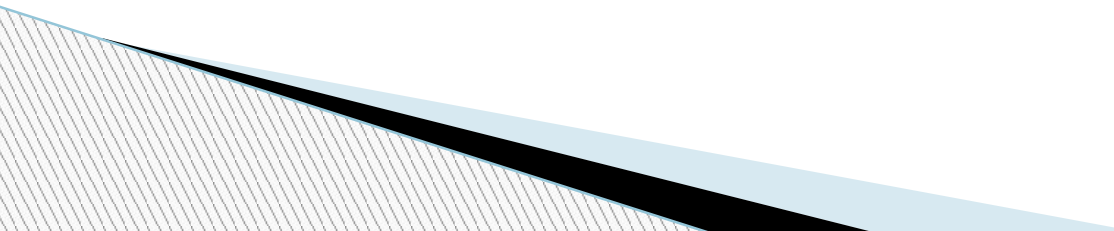
Context is always important

c.f. for a student the mark required for a pass is less than the mark required to be admitted to a postgraduate degree



Example 1: Cantor WJ, 2009 NEJM

Hypothesis of the study :Patients with MI treated with fibrinolysis and transferred for early angioplasty (PCI) have less complications vs patients treated with standard therapy.

- Study design: randomized non-blinded multicentre trial (experimental)
 - Primary endpoint: combined death, reinfarction, recurrent ischaemia, new or worsening heart failure or cardiogenic shock at 30days
 - Secondary endpoint: complications(bleeding)
- 

Example 1: Cantor WJ, 2009 NEJM

- Sample: event rate of primary endpoint: 21%, 5% lost to follow-up, power of 80%, and alpha level of 0.05
- Relative risk reduction of 30% in the early PCI group: $n=1200$

Results

End points	Standard Treatment (n=522)	Early PCI (n=536)	Relative Risk (95%CI)	P-value
Primary end point composite	90 (17%)	59(11%)	0.64(0.47-0.87)	0.004
Death	18 (3.5%)	24(4.5%)	1.30 (0.71-2.36)	0.39
Heart failure	29(5.6%)			
Bleeding	84 (16.1%)	110 (20.5%)	1.27 (0.98-1.65)	0.06

Example 1

Conclusions: The primary end point a composite of death, congestive heart failure occurred less frequently with early PCI with standard therapy

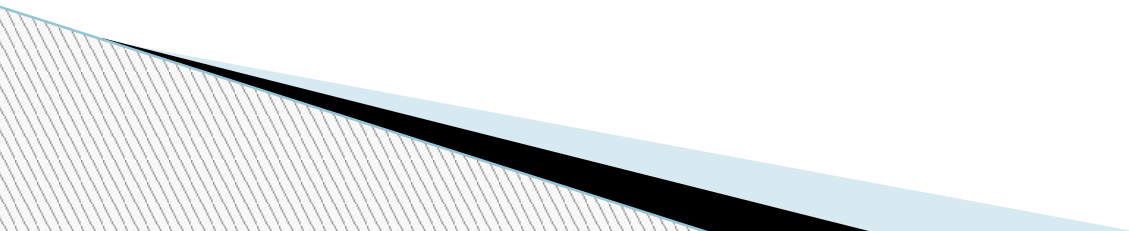
No significant differences in the bleeding between both groups, **but look at CI !!!!!!!**

It was statistically significant and clinically relevant, meaning it can be used in the clinical setting!



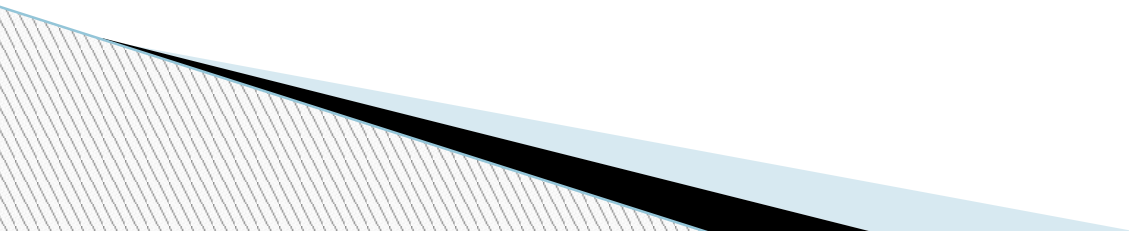
Example 1

Context is always important e.g. clinical implications of bleeding compared to composite endpoint



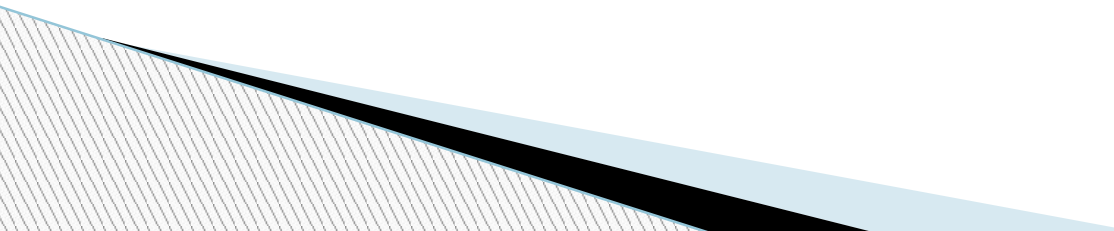
Example 2: Thai HIV vaccine trial RV144

Vaccination with ALVAC and AIDSVAX to prevent HIV-1 infection in Thailand NEJM 2009 361: 2209-2220



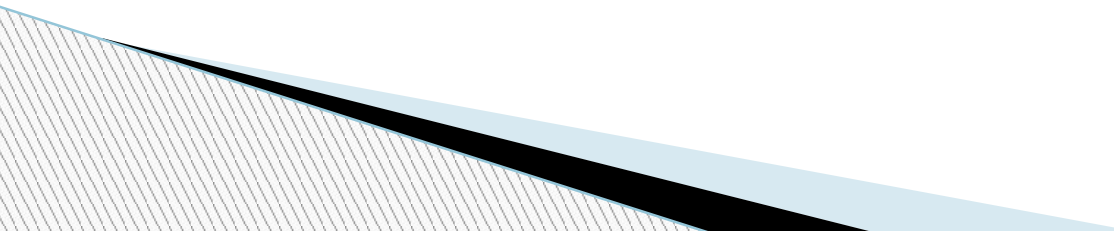
Example 2: Thai HIV vaccine trial RV144

In the ITT analysis involving 16,402 participants there was a trend towards the prevention of HIV-1 infection among vaccine recipients with a vaccine efficacy of 26.4% (95% CI -4.0 to 47.9; $P=0.08$).



Example 2: Thai HIV vaccine trial RV144

In the mITT analysis involving 16,395 participants (excluding 7 participants who were found to have had HIV-1 infection at baseline) vaccine efficacy was 31.2% (95% CI 1.1 to 52.1 ; $P=0.04$)

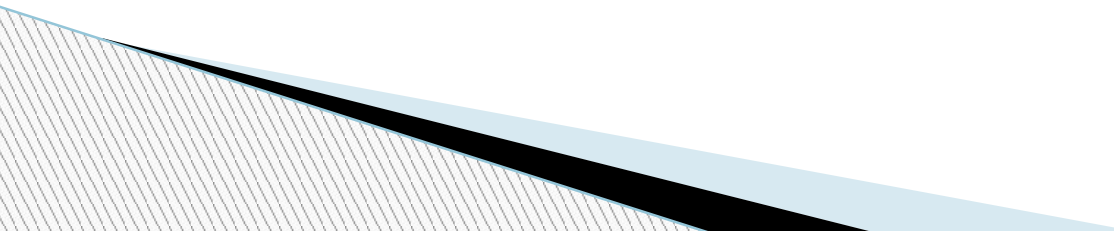


Example 2: Thai HIV vaccine trial RV144

The ALVAC-HIV and AIDSVAX B/E vaccine regimen may reduce the risk of HIV infection in a community based population with largely heterosexual risk

Although results show only a modest benefit they offer insight for future research

NB – context is important – IAVI have been looking for an HIV vaccine since 1995

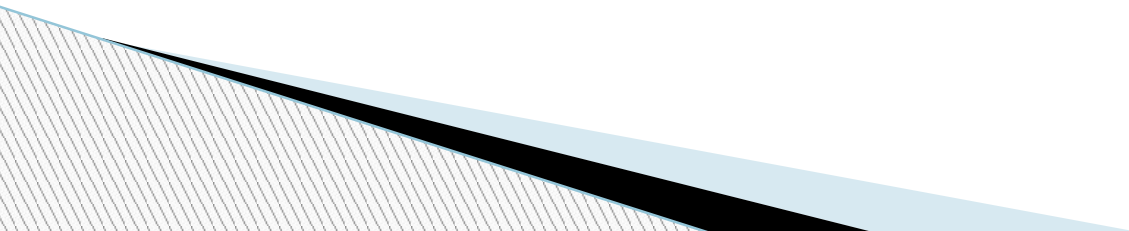


Conclusions

Non-significant does not mean no effect

We should not only rely on the p- value

“Statistically significant” does not always mean
“clinically significant”

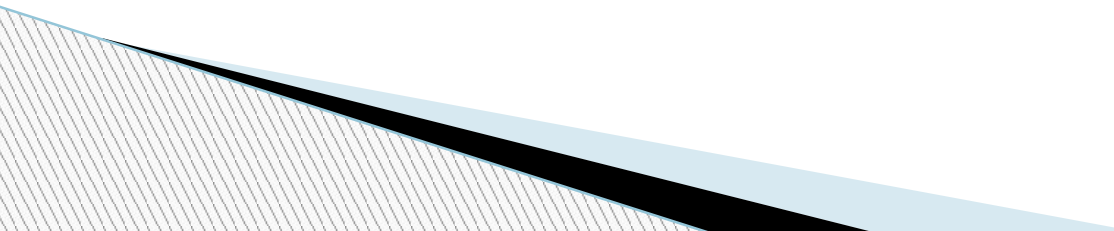


Conclusions

We have to look at the broader picture and consider results in the light of study design, sample size calculations, 95%CI, limitations, and current scientific knowledge

Very often further work is required

If a finding is unique, but you cannot explain it ,
continue researching!!!!



Thank You

