Power & Sample Size

Dr. Andrea Benedetti

Plan

- Review of hypothesis testing
- Power and sample size
  - Basic concepts
  - Formulae for common study designs
- Using the software

When should you think about power & sample size?

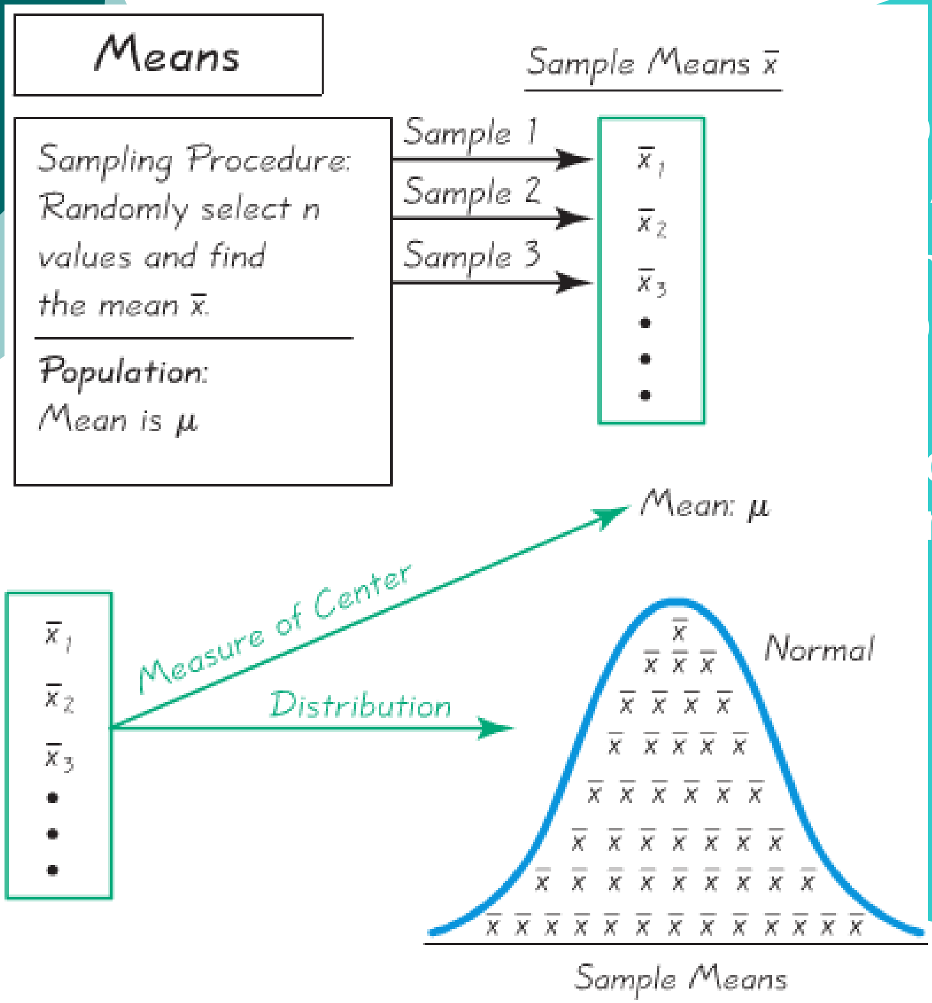
- You should start thinking about statistics when you are planning your study
- Often helpful to consult a statistician at this stage…
- You should also perform a power calculation at this stage

REVIEW

Basic concepts

- Descriptive statistics
  - Raw data → graphs, averages, variances, categories
- Inferential statistics
  - Raw data → Summary data → Draw conclusions about a population from a sample

Suppose we are interested in the height of men...



pulation= All men

mpossible measure all men

opulation mean=μ

Study sample 2

Mean Ht

Study sample

Mean Ht

inference

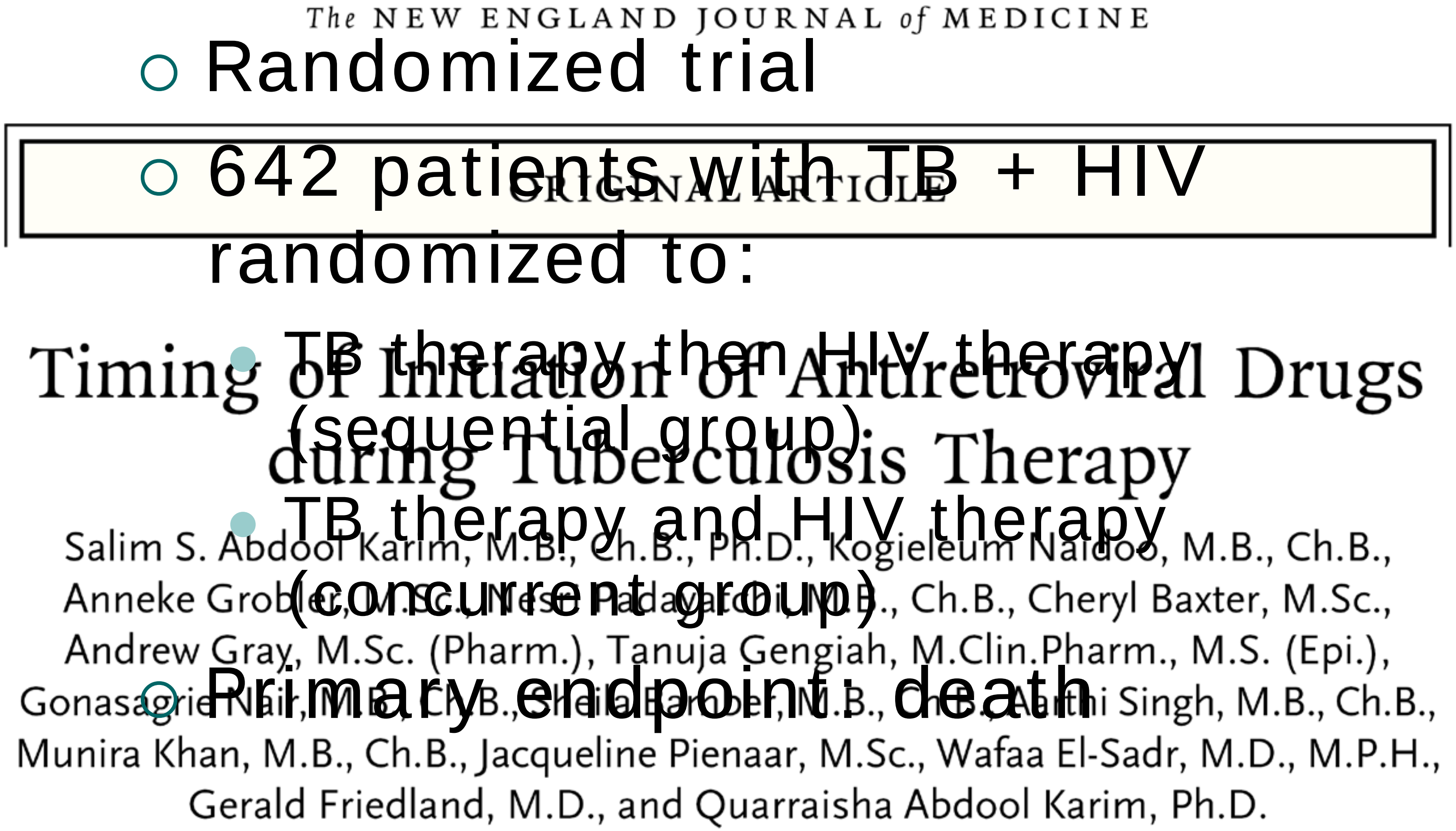
Mean ht = “statistic”

Populations --- samples

- Suppose, we are interested in the mean height of men.
  - we have a population "men" and a parameter of interest, their height
  - a hypothetical population that includes all men
- it is impossible to survey/measure the entire population
- Instead, take a subset of this population (“sample”)
  - use this sample to draw inferences about the population, given conditions
    - measure the mean height of men (“a statistic”) in the sample
    - draw inferences about the parameter of interest in the population
  - “inference” because there is uncertainty involved in drawing conclusions about the population based upon a sample
- the sample we get is one of a large number of potential samples
  - the statistic in question (mean height) varies from sample to sample
  - it has a distribution called a sampling distribution
  - this distribution is used to understand the uncertainty in our estimate of the population parameter

An example

- ○ Randomized trial
- ○ 642 patients with TB + HIV randomized to:
  - ● TB therapy then HIV therapy (sequential group)
  - ● TB therapy and HIV therapy (concurrent group)
- ○ Primary endpoint: death



The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Timing of Initiation of Antiretroviral Drugs during Tuberculosis Therapy

Salim S. Abdool Karim, M.B., Ch.B., Ph.D., Kogieleum Naidoo, M.B., Ch.B., Anneke Grobler, M.Sc., Nesri Padayatchi, M.B., Ch.B., Cheryl Baxter, M.Sc., Andrew Gray, M.Sc. (Pharm.), Tanuja Gengiah, M.Clin.Pharm., M.S. (Epi.), Gonasagrie Nair, M.B., Ch.B., Sheila Bamber, M.B., Ch.B., Aarthi Singh, M.B., Ch.B., Munira Khan, M.B., Ch.B., Jacqueline Pienaar, M.Sc., Wafaa El-Sadr, M.D., M.P.H., Gerald Friedland, M.D., and Quarraisha Abdool Karim, Ph.D.

Hypothesis Test…1

- Setting up and testing hypotheses is an essential part of statistical inference
  - usually some theory has been put forward
    - e.g. claiming that a new drug is better than the current drug for treatment of the same illness
    - Does the concurrent group have less risk of death than the sequential group?

Hypothesis Test… 2

- The question of interest is simplified into two competing hypotheses between which we have a choice
  - H0: the null hypothesis
  - HA: the alternative hypotheis
- These two competing hypotheses are not treated on an equal basis
  - special consideration is given to H0
  - if one of the hypotheses is 'simpler' we give it priority
  - a more 'complicated' theory is not adopted unless there is sufficient evidence against the simpler one

Hypothesis Test… 3

- The outcome of a hypothesis test:
  - final conclusion is given in terms of H0.
    - "Reject H0 in favour of HA”
    - "Do not reject H0";
    - we *never* conclude "Reject HA", or even "Accept HA"
- If we conclude "Do not reject H0", this does not necessarily mean that H0 is true, it only suggests that there is not sufficient evidence against H0 in favour of HA.
  - Rejecting H0 suggests that HA *may* be true.

TYPE I AND TYPE II ERRORS

Type I and II errors

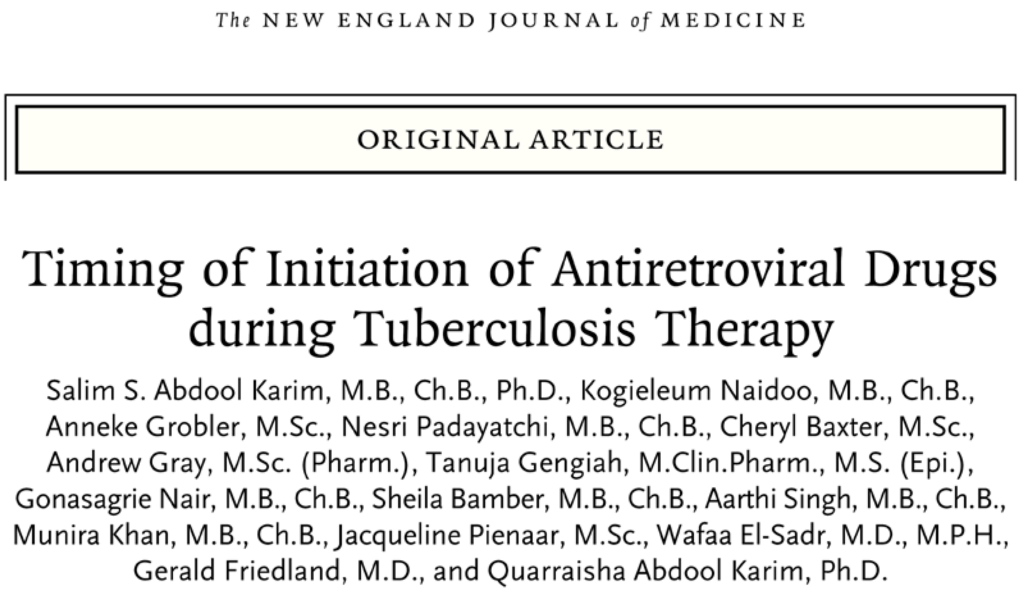
Truth

We don't know if H0 is true!!!

	H_0 is true	H_A is true
Reject H_0		
Do not reject H_0		

Result of the test

An example



The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Timing of Initiation of Antiretroviral Drugs during Tuberculosis Therapy

Salim S. Abdool Karim, M.B., Ch.B., Ph.D., Kogieleum Naidoo, M.B., Ch.B., Anneke Grobler, M.Sc., Nesri Padayatchi, M.B., Ch.B., Cheryl Baxter, M.Sc., Andrew Gray, M.Sc. (Pharm.), Tanuja Gengiah, M.Clin.Pharm., M.S. (Epi.), Gonasagrie Nair, M.B., Ch.B., Sheila Bamber, M.B., Ch.B., Aarthi Singh, M.B., Ch.B., Munira Khan, M.B., Ch.B., Jacqueline Pienaar, M.Sc., Wafaa El-Sadr, M.D., M.P.H., Gerald Friedland, M.D., and Quarraisha Abdool Karim, Ph.D.

- Randomized trial
- 642 patients with TB + HIV randomized to:
  - TB therapy then HIV therapy (sequential group)
  - TB therapy and HIV therapy (concurrent group)
- Primary endpoint: death

Example

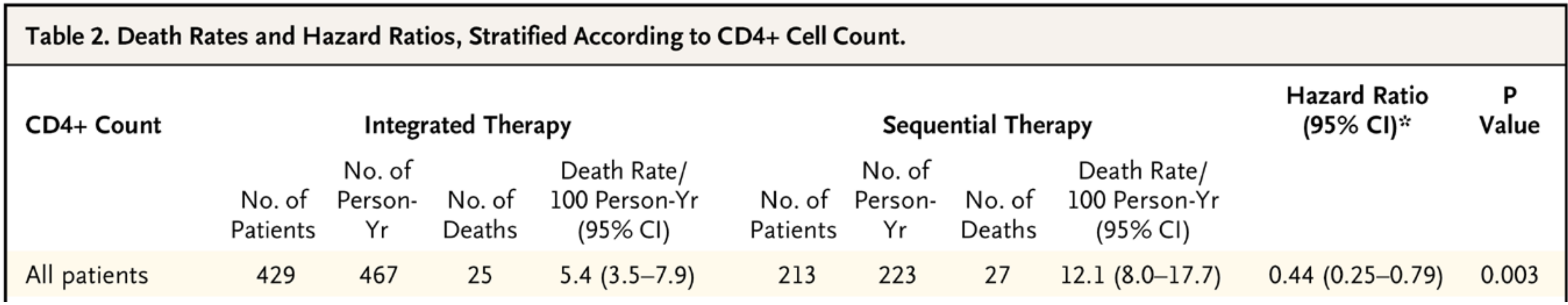


Table 2. Death Rates and Hazard Ratios, Stratified According to CD4+ Cell Count.

CD4+ Count	Integrated Therapy				Sequential Therapy				Hazard Ratio (95% CI)*	P Value
	No. of Patients	No. of Person-Yr	No. of Deaths	Death Rate/ 100 Person-Yr (95% CI)	No. of Patients	No. of Person-Yr	No. of Deaths	Death Rate/ 100 Person-Yr (95% CI)		
All patients	429	467	25	5.4 (3.5–7.9)	213	223	27	12.1 (8.0–17.7)	0.44 (0.25–0.79)	0.003

- What was H0?
- What was HA?
- What was the outcome?

Example

- H0: the death rate is the same in the two groups
- HA: the death rate is different in the two groups

	H_0 is true	H_A is true
Reject H_0		
Do not reject H_0		

α=Type I error

- In repeated sampling, this test will commit a type I error 100*α% of the time. We control this by selecting the significance level of our test (α).

Type I and Type II errors

- more concerned about Type I error
  - concluding that there is a difference when there really is no difference than Type II errors
- So... set Type I error at 0.05
  - then choose the procedure that minimizes Type II error (or equivalently, maximizes power)

Type I and Type II errors

- If we do not reject H0, we are in danger of committing a type II error.
  - i.e. The means are different, but we did not see it.
- If we do reject H0, we are in danger of committing a type I error.
  - i.e. the means are not truly different, but we have declared them to be different.

Going back to our example

- What is a type I error?
  - Rejecting H0 when H0 is true
  - Concluding that the concurrent group has a different death rate than the sequential group, when there truly is no difference.
- What is a type II error?
  - Not rejecting H0 when there really is a difference.

Power

- Power is the probability of rejecting the H0 when HA is true.
- You should design your study to have enough subjects to detect important effects, but not too many
- We usually aim for power ≥ 80%

Clinical significance vs. statistical significance

		Clinical significance	
		Yes	No
Statistical Significance	Yes		
	No		

WHAT AFFECTS POWER?

Credit for many slides: Juli Atherton, PhD

Generic ideas about sample size

For a two sided test with type I error=α to have at least 100*(1-β)% power against a nonzero difference Δ then:

$$Z_{\alpha/2}SE_{null}+Z_{\beta}SE_{alt}<\Delta$$

We can simplify this to:

$$(Z_{\alpha/2}+Z_{\beta})*SE<\Delta$$

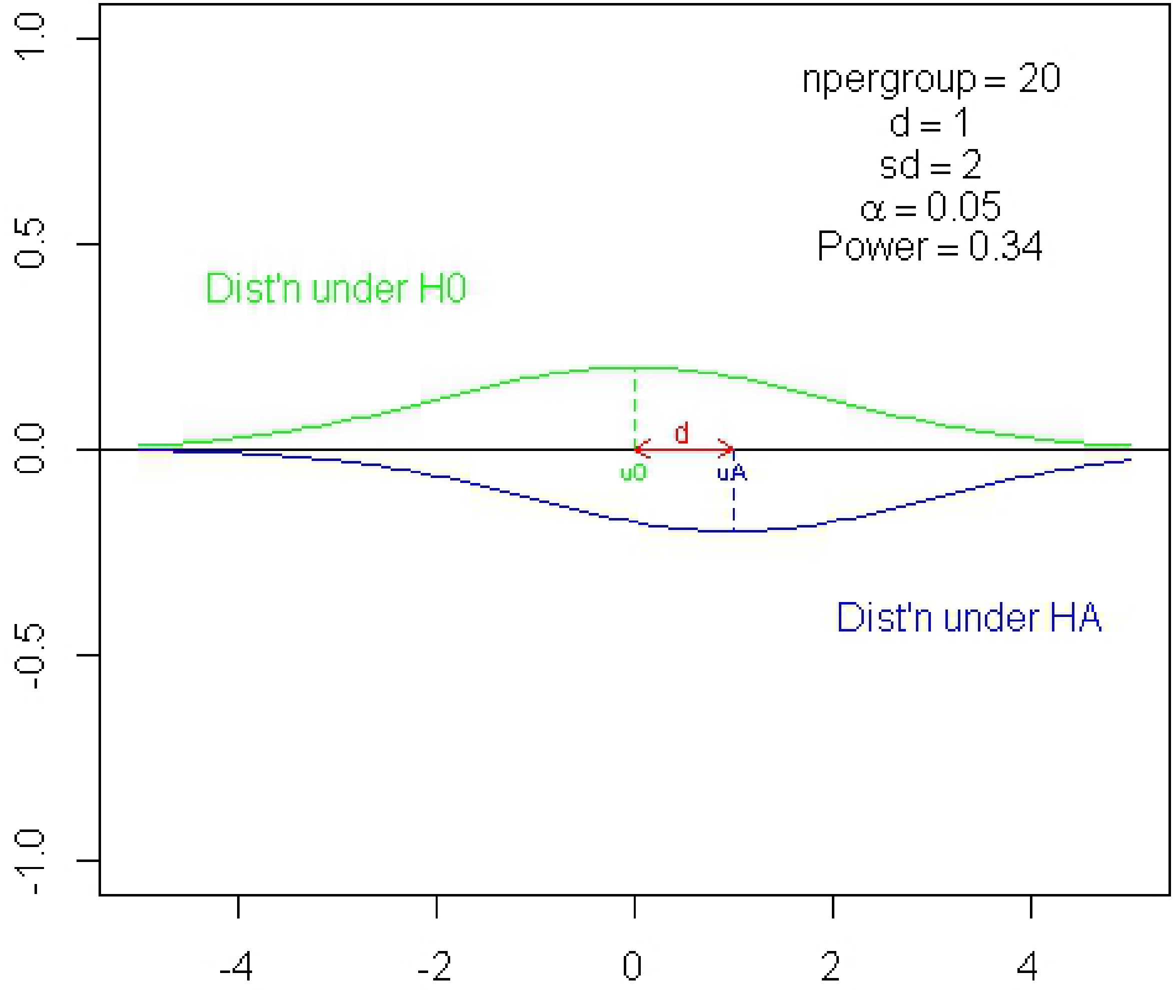
Or more:

$$(Z_{\alpha/2}+Z_{\beta})^2*Var(Parameter\ estimate)<\Delta^2$$

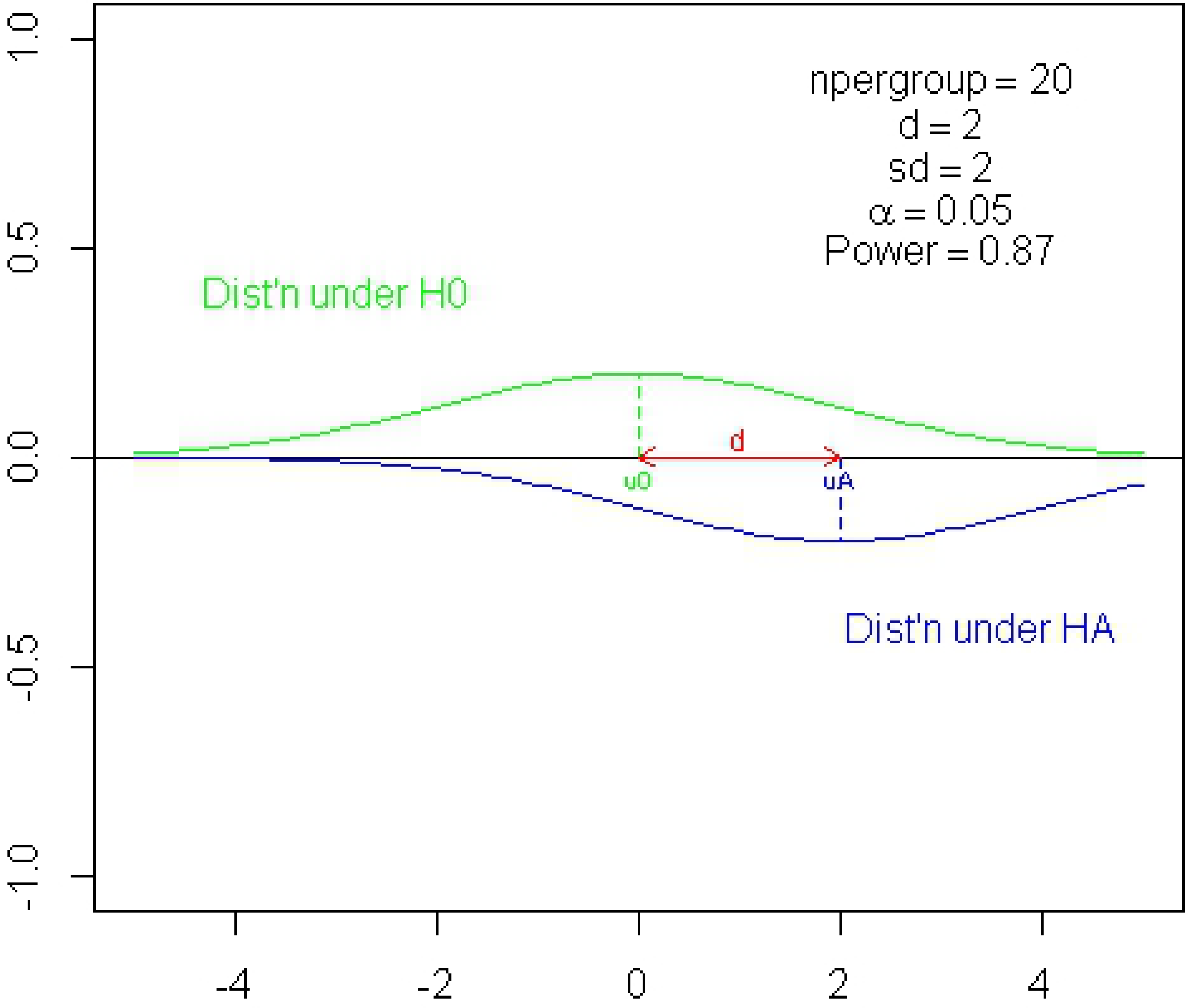
What info do we need to compute power?

- Type I error rate (α)
- The sample size
- The detectable difference
- The variance of the measure
  - will depend on the type of outcome

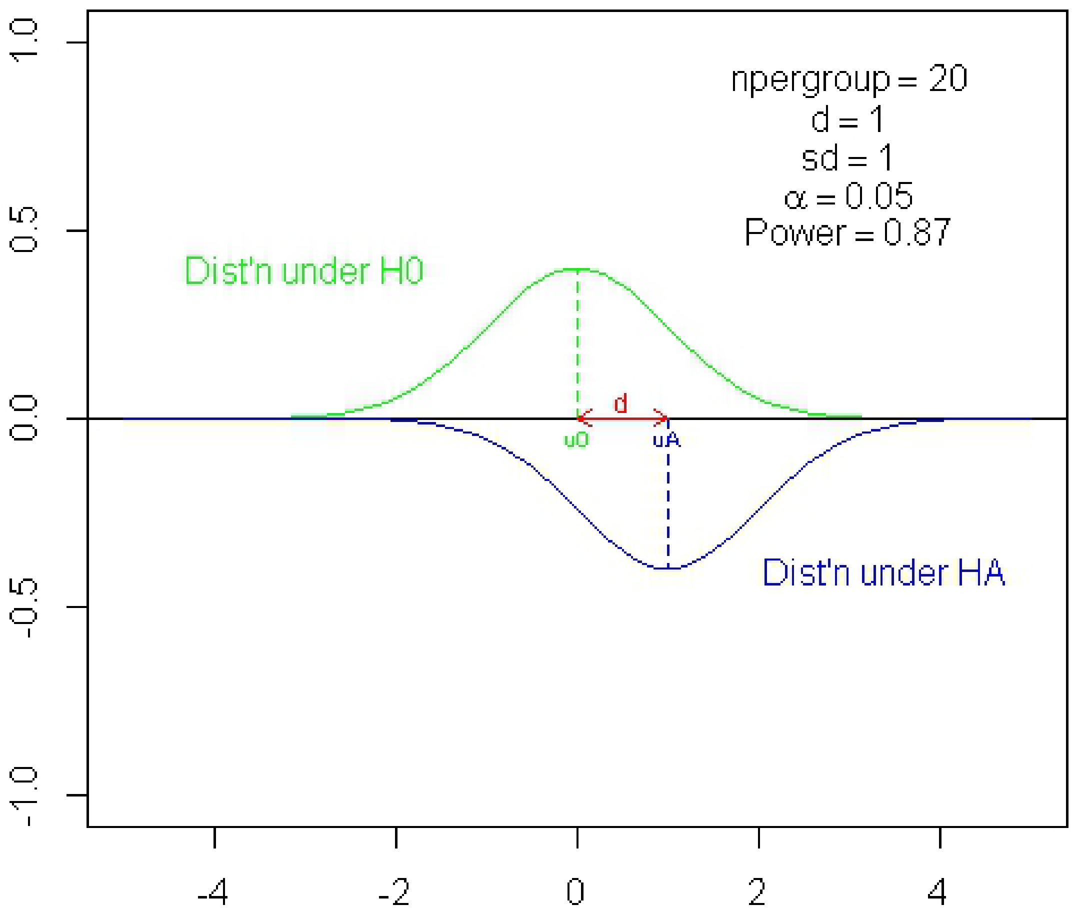
What affects power?



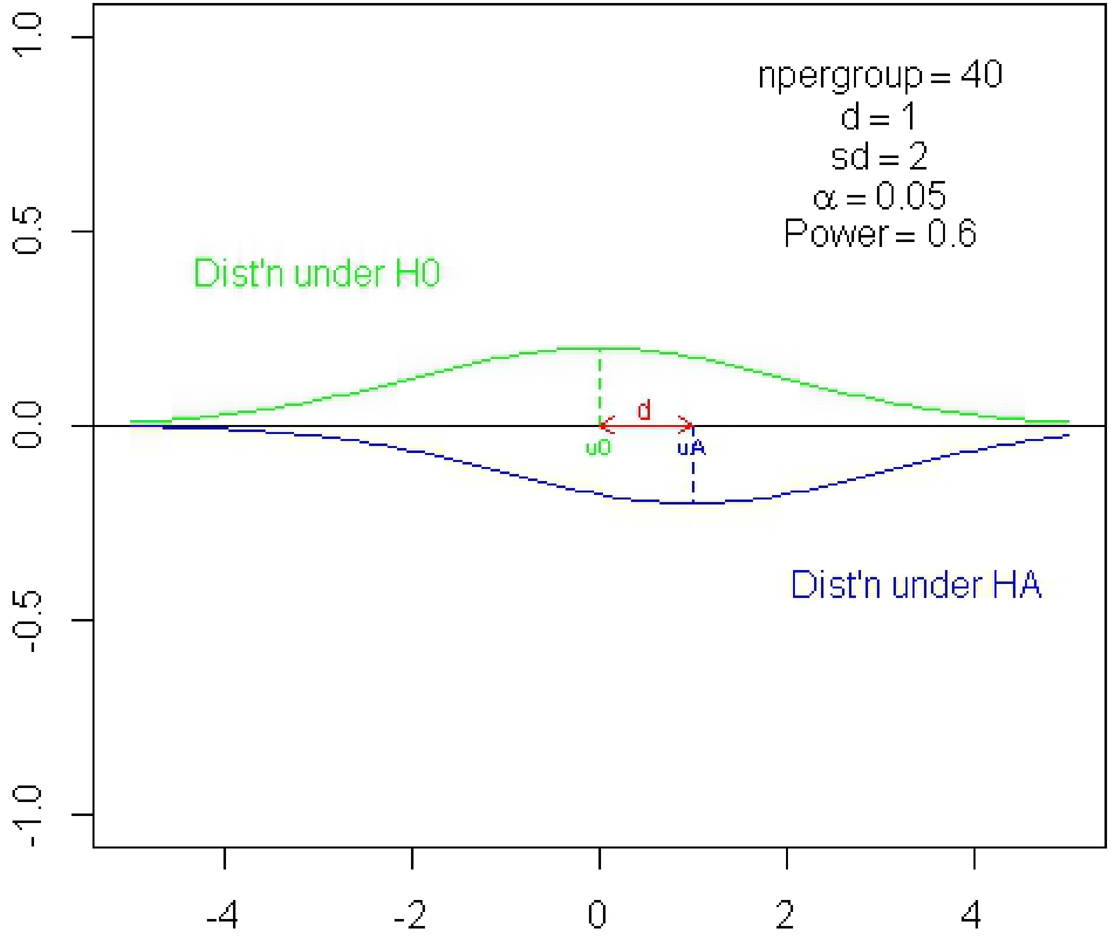
What happens if we increase the detectable difference?



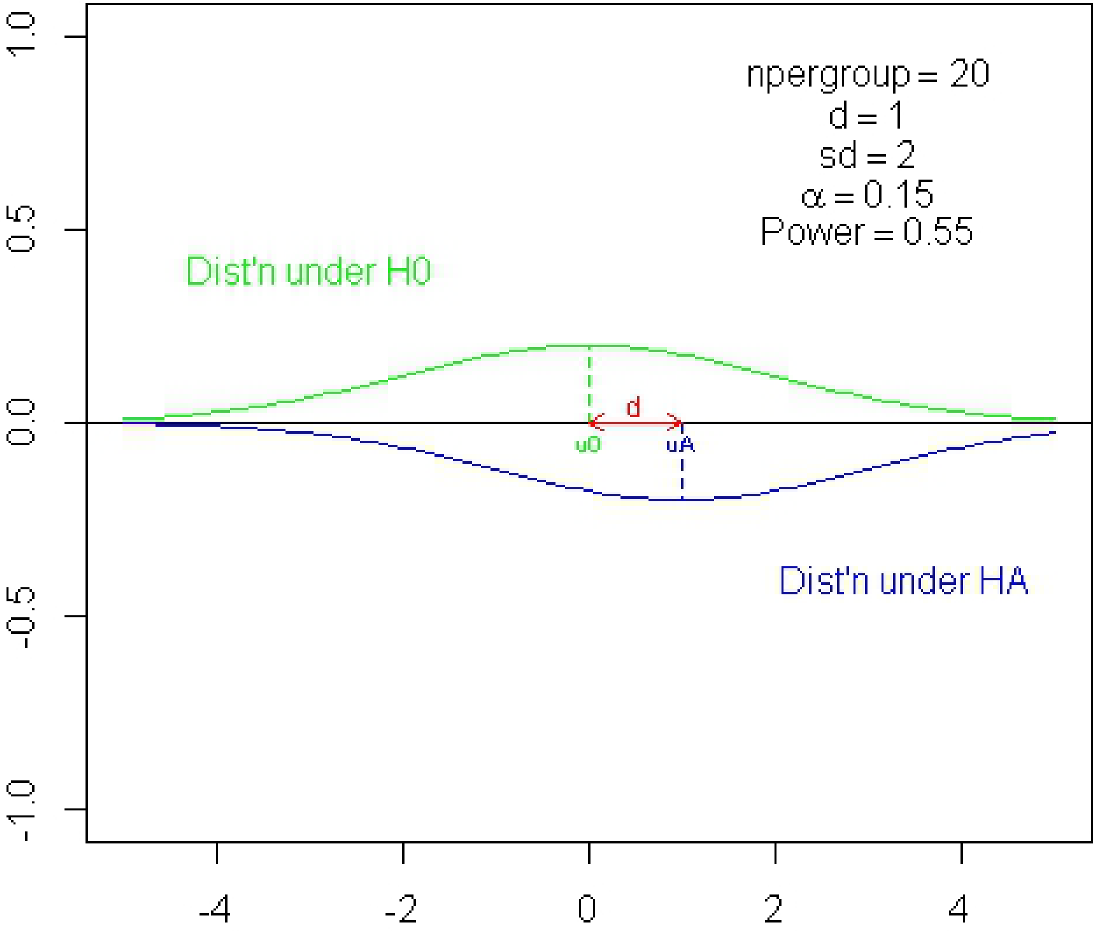
What happens if the sd decreases?



What happens if n increases?



What happens if we increase α?



So, what affects power?

- Size of the detectable effect
- Number of subjects
- Variance of the measure
- Level of significance
  - http://bcs.whfreeman.com/ips4e/pages/bcs-main.asp?v=category&s=00010&n=99000&i=99010.01&o=

SAMPLE SIZE CALCULATIONS

Credit for many slides: Juli
Atherton, PhD

Binary outcomes

- Objective: to determine if there is evidence of a statistical difference in the comparison of interest between two regimens (A and B)

- H0: The two treatments are not different ($\pi_A = \pi_B$)
- HA: The two treatmetns are different ($\pi_A \neq \pi_B$)

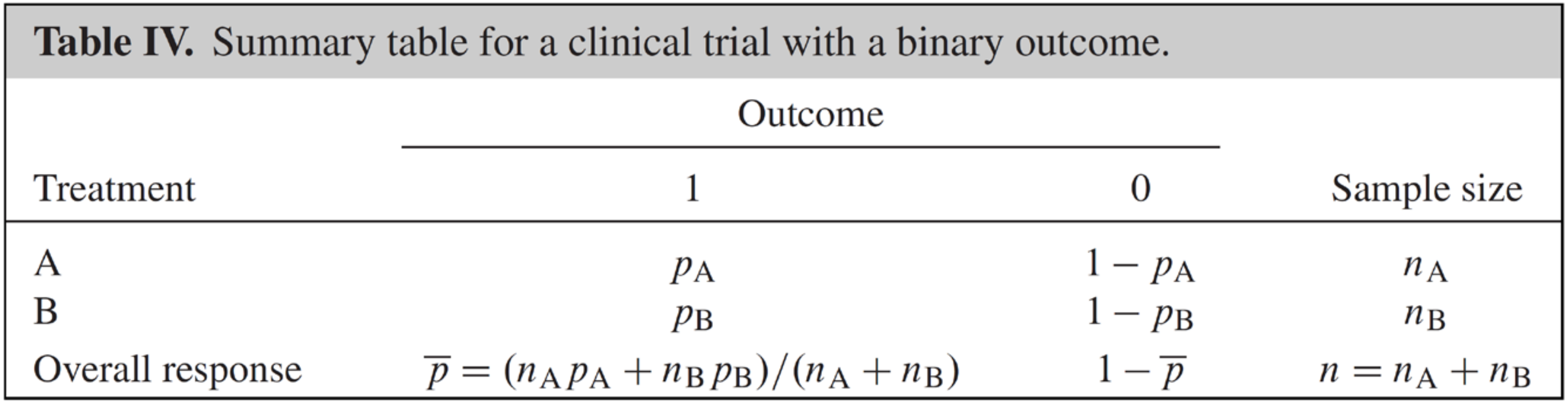


Table IV. Summary table for a clinical trial with a binary outcome.

	Outcome		
Treatment	1	0	Sample size
A	p_A	$1 - p_A$	n_A
B	p_B	$1 - p_B$	n_B
Overall response	$\overline{p} = (n_A p_A + n_B p_B)/(n_A + n_B)$	$1 - \overline{p}$	$n = n_A + n_B$

Sample size for binary, parallel arm superiority trial:

- Equally sized arms:

$$n_{\mathrm{A}} = \frac{\left[Z_{1-\beta} + Z_{1-\alpha/2}\right]^2 (\pi_{\mathrm{A}}(1-\pi_{\mathrm{A}}) + \pi_{\mathrm{B}}(1-\pi_{\mathrm{B}}))}{(\pi_{\mathrm{A}} - \pi_{\mathrm{B}})^2}$$

- 90% power:

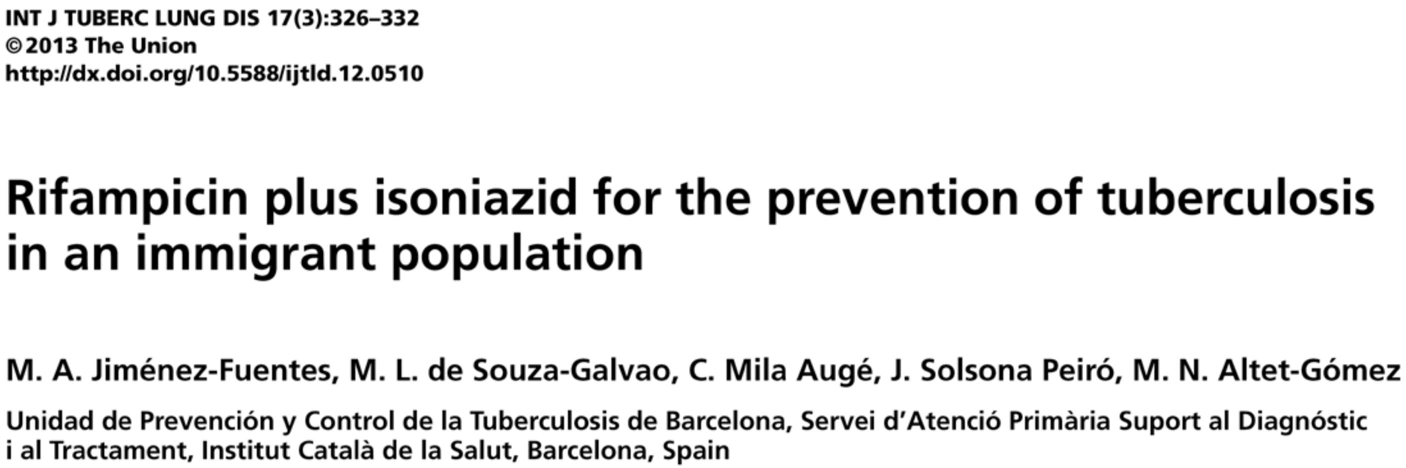
$$n_{\mathrm{A}} = \frac{5.25}{(\pi_{\mathrm{A}} - \pi_{\mathrm{B}})^2}.$$

- 80% power:

$$n_{\mathrm{A}} = \frac{4}{(\pi_{\mathrm{A}} - \pi_{\mathrm{B}})^2}$$

Stat Med 2012 for details

Example



INT J TUBERC LUNG DIS 17(3):326–332
©2013 The Union
http://dx.doi.org/10.5588/ijtld.12.0510

Rifampicin plus isoniazid for the prevention of tuberculosis in an immigrant population

M. A. Jiménez-Fuentes, M. L. de Souza-Galvao, C. Mila Augé, J. Solsona Peiró, M. N. Altet-Gómez

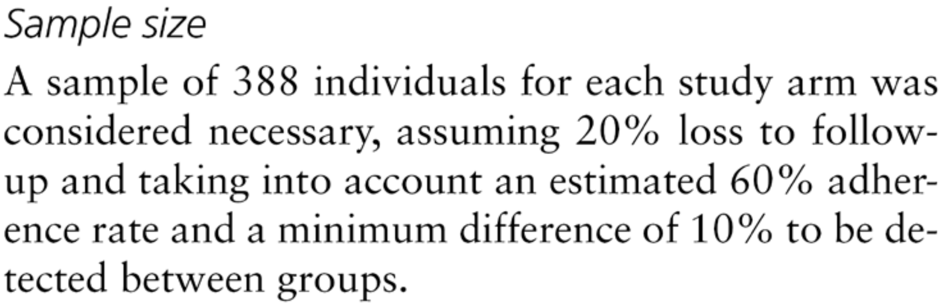
Unidad de Prevención y Control de la Tuberculosis de Barcelona, Servei d'Atenció Primària Suport al Diagnóstic i al Tractament, Institut Català de la Salut, Barcelona, Spain

ns

s.

- ○ π_{3M}=proportion that adhered to treatment in the 3 month group
- ○ π_{6M}=proportion that adhered in the 6 month group

- What info do we need to calculate sample size?
  - Desired power
  - Detectable difference
  - Proportion in the ‘control’ arm
  - Type I error



Sample size

A sample of 388 individuals for each study arm was considered necessary, assuming 20% loss to follow-up and taking into account an estimated 60% adherence rate and a minimum difference of 10% to be detected between groups.

This is terrible – what info is this missing??

$$n_{\mathrm{A}} = \frac{\left[Z_{1-\beta} + Z_{1-\alpha/2}\right]^2 \left(\pi_{\mathrm{A}}(1-\pi_{\mathrm{A}}) + \pi_{\mathrm{B}}(1-\pi_{\mathrm{B}})\right)}{\left(\pi_{\mathrm{A}} - \pi_{\mathrm{B}}\right)^2}$$

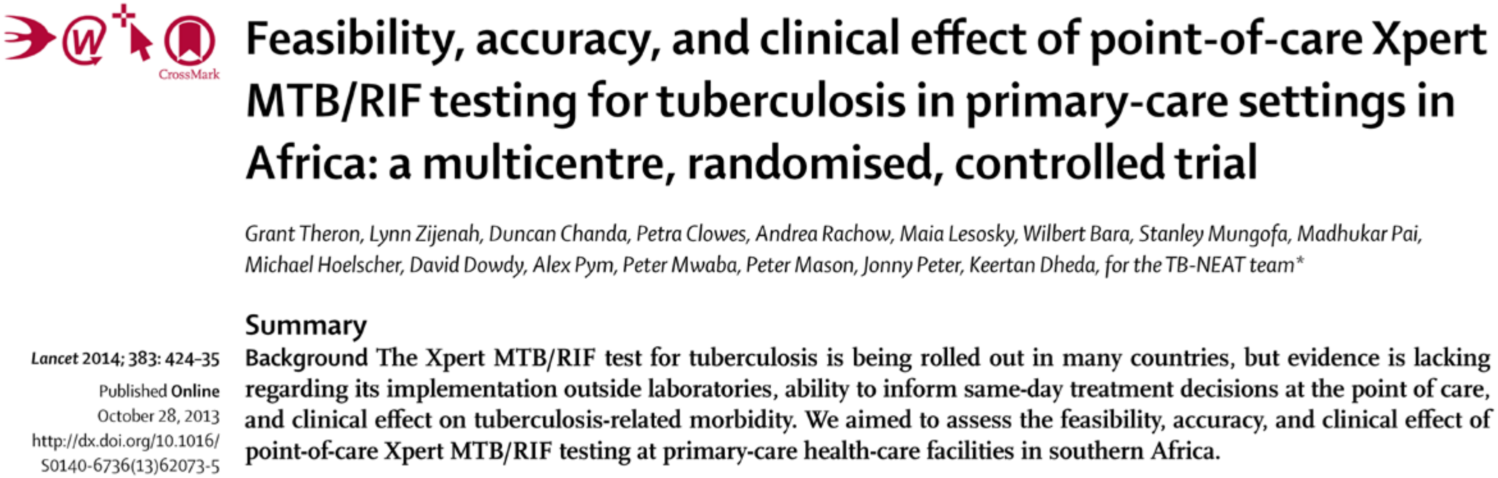
- 80% power:
- n_A $=[(1.96+0.84)^2*(0.6*0.4+0.7*0.3)]/[(0.1)^2]$
  =353
  =with 10% loss to follow up: 1.1*353=388 PER GROUP

- To achieve 80% power, a sample of 388 individuals for each study arm was considered necessary, assuming 10% loss to follow up and taking into account an estimated 60% adherence rate and a minimum difference of 10% to be detected between groups with alpha=.05.

What about continuous data?

- $n_{pergroup} = [2*(Z_{1-\alpha/2} + Z_{1-\beta})^2 * Var(Y)]/\Delta^2$
- So what info do we need?
  - Desired power
  - Type I error
  - Variance of outcome measure
  - Detectable difference

Continuous Data example



Feasibility, accuracy, and clinical effect of point-of-care Xpert MTB/RIF testing for tuberculosis in primary-care settings in Africa: a multicentre, randomised, controlled trial

*Grant Theron, Lynn Zijenah, Duncan Chanda, Petra Clowes, Andrea Rachow, Maia Lesosky, Wilbert Bara, Stanley Mungofa, Madhukar Pai, Michael Hoelscher, David Dowdy, Alex Pym, Peter Mwaba, Peter Mason, Jonny Peter, Keertan Dheda, for the TB-NEAT team**

Lancet 2014; 383: 424–35
Published **Online** October 28, 2013
http://dx.doi.org/10.1016/S0140-6736(13)62073-5

Summary

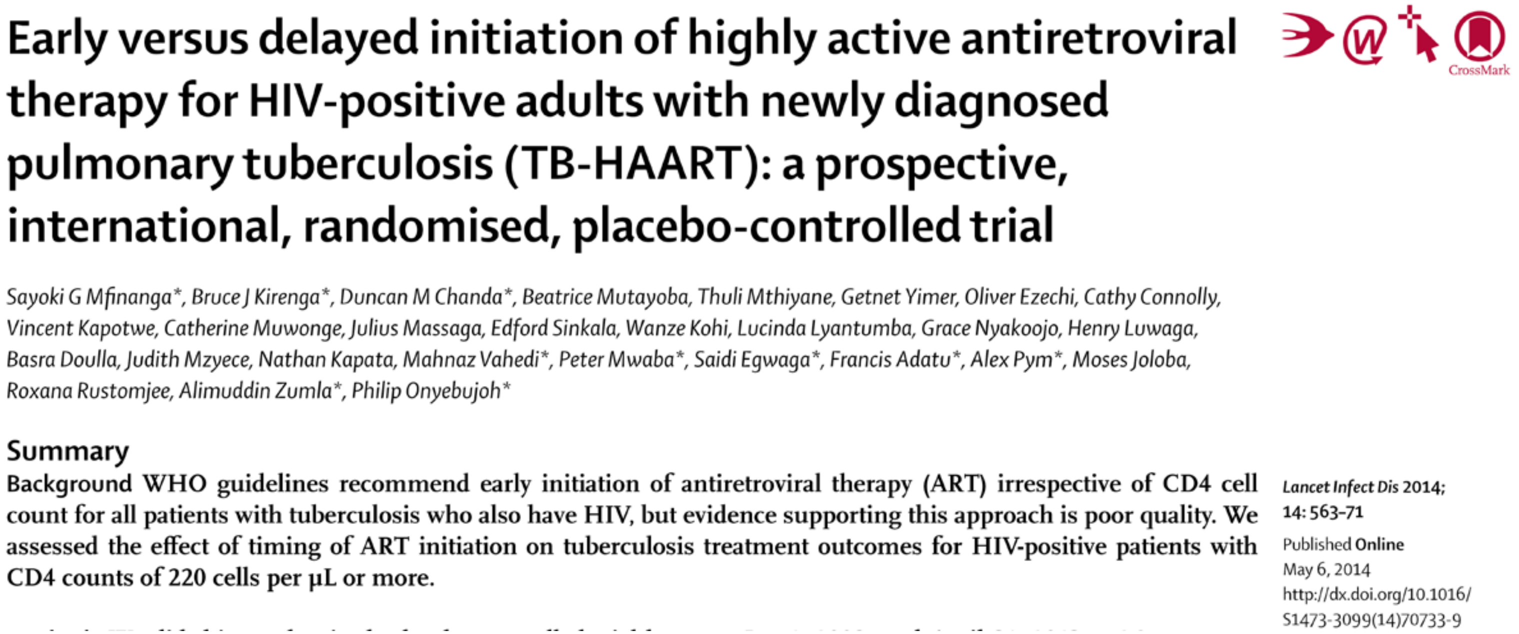
Background The Xpert MTB/RIF test for tuberculosis is being rolled out in many countries, but evidence is lacking regarding its implementation outside laboratories, ability to inform same-day treatment decisions at the point of care, and clinical effect on tuberculosis-related morbidity. We aimed to assess the feasibility, accuracy, and clinical effect of point-of-care Xpert MTB/RIF testing at primary-care health-care facilities in southern Africa.

Our primary outcome was tuberculosis-related morbidity (graded using the TBscore and Karnofsky performance score (see appendix for definitions)

From the appendix:

- We projected the difference in TBscore between arms to be 1 (the minimally important clinical difference).
- We assumed, based on previous studies, that the within group standard deviation would be 2 points in each arm.
- With an alpha value of 5% (two-sided) and a desired power of 80%, and assuming equal numbers in each arm, we required approximately 63 culture-positive patients in each arm.
- To account for deaths, loss to follow-up, withdrawals, and missing data, we inflated this by 30% (~82 culture-positive).

Example: Time to event (survival)



Early versus delayed initiation of highly active antiretroviral therapy for HIV-positive adults with newly diagnosed pulmonary tuberculosis (TB-HAART): a prospective, international, randomised, placebo-controlled trial

Sayoki G Mfinanga, Bruce J Kirenga*, Duncan M Chanda*, Beatrice Mutayoba, Thuli Mthiyane, Getnet Yimer, Oliver Ezechi, Cathy Connolly, Vincent Kapotwe, Catherine Muwonge, Julius Massaga, Edford Sinkala, Wanze Kohi, Lucinda Lyantumba, Grace Nyakoojo, Henry Luwaga, Basra Doulla, Judith Mzyece, Nathan Kapata, Mahnaz Vahedi*, Peter Mwaba*, Saidi Egwaga*, Francis Adatu*, Alex Pym*, Moses Joloba, Roxana Rustomjee, Alimuddin Zumla*, Philip Onyebujoh**

Summary

Background WHO guidelines recommend early initiation of antiretroviral therapy (ART) irrespective of CD4 cell count for all patients with tuberculosis who also have HIV, but evidence supporting this approach is poor quality. We assessed the effect of timing of ART initiation on tuberculosis treatment outcomes for HIV-positive patients with CD4 counts of 220 cells per μL or more.

Lancet Infect Dis 2014;
14: 563–71

Published **Online**
May 6, 2014
http://dx.doi.org/10.1016/
S1473-3099(14)70733-9

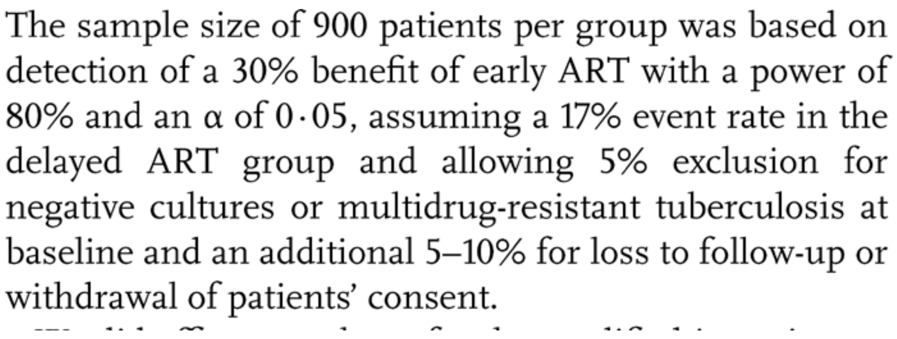
- Primary endpoint: death, failure of TB treatment, recurrence TB at 12 months
- Follow up time: 24 months

“Rule of thumb”

$$n = \frac{4}{T(\sqrt{\theta_0} - \sqrt{\theta_1})^2} \,.$$

- n per group
- θ_0=death rate per unit time in control group
- θ_1=death rate per unit time in the experimental group
- T=follow up time
- With alpha=.05 and power=80%

- θ_0=death rate per unit time in control group=0.17/24
- θ_1=death rate per unit time in the experimental group=0.12/24
- T=follow up time=24 months



The sample size of 900 patients per group was based on detection of a 30% benefit of early ART with a power of 80% and an α of 0·05, assuming a 17% event rate in the delayed ART group and allowing 5% exclusion for negative cultures or multidrug-resistant tuberculosis at baseline and an additional 5–10% for loss to follow-up or withdrawal of patients' consent.

Power & Sample Size

EXTENSIONS & THOUGHTS

Accounting for losses to follow up

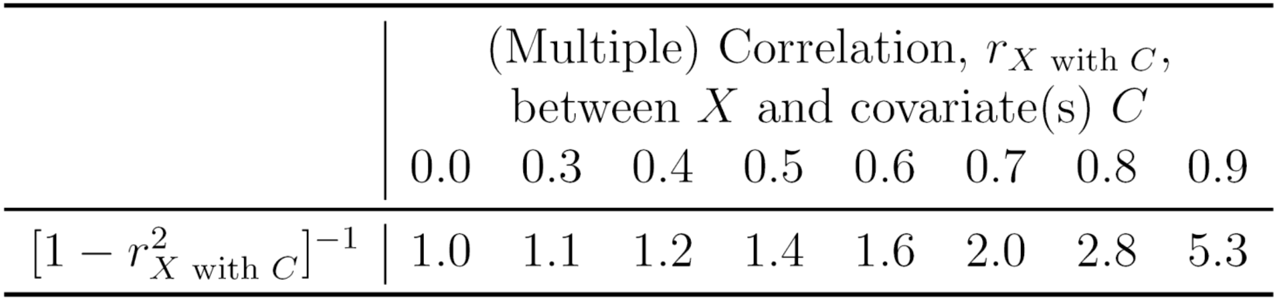
- Whatever sample size you arrive at, inflate it to account for losses to follow up…

Adjusting for confounders

- ...will be necessary if the data come from observational study, or in some cases in clinical trials (more tomorrow)
- Rough rule of thumb: need about 10 observations or events per variable you want to include

Accounting for confounders

- If you want to be fancier:



	(Multiple) Correlation, $r_{X \text{ with } C}$, between X and covariate(s) C							
	0.0	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$[1 - r^2_{X \text{ with } C}]^{-1}$	1.0	1.1	1.2	1.4	1.6	2.0	2.8	5.3

Table 5: Sample size multiplier to cover the cost of adjusting for one or more confounding variables.The multiplier, $[1\text{-}r^2_{X \text{ with } C}]^{-1}$ is given as a function of the (multiple) correlation, $r_{X \text{ with } C}$, between *X* and the set of confounding variables *C*. It is rounded up, and given to two significant digits.

Adjusting for multiple testing

- Lots of opinion about whether you should do this or not
- Simplest case: Bonferroni adjustment
  - use α/n instead of α
  - over-conservative
  - reduces power a lot!
- Many other alternatives – false discovery rate is a good one

Non-independent samples

- before-after measurements
- multiple measurements from the same person over time
- measurements from subjects from the same family/household
- Geographic groupings

- These measurements are likely to be correlated.
- This correlation must be taken into account or inference may be wrong!
    - p-values may be too small, confidence intervals too tight

Clustered data

- Easy to account for this via "the design effect"
- Use standard sample size then inflate by multiplying by the design effect:
- Deff=1+(m-1)ρ
  - m=average cluster size
  - ρ=intraclass correlation coefficient
    - A measure of correlation between subjects in the same cluster

Subgroups

- If you want to detect effects in subgroups, then you should consider this in your sample size calculations.

Where to get the information?

- All calculations require some information
  - Literature
    - similar population?
    - same measure?
  - Pilot data
- Often best to present a table with a range of plausible values and choose a combination that results in a conservative (i.e. big) sample size

Wrapping up

- You should specify all the ingredients as well as the sample size
- We have focused here on estimating sample size for a desired power, but could also estimate power for a given sample size