Statistical issues in randomized trials

Matthew Law | 28 August 2021

Contents

- Hypothesis testing & confidence intervals
- Power - small trials
- Randomisation and intention to treat
- Subgroup analyses
- Non-inferiority / equivalence trials

Contents

- Hypothesis testing & confidence intervals
- Power - small trials
- Randomisation and intention to treat
- Subgroup analyses
- Non-inferiority / equivalence trials

Framework to fix ideas

- Two arm randomized trial
  - X patients randomized to each of treatments A and B
- Treatments A and B compared using key endpoints
  - Survival
  - Proportions detectable HIV viral load
  - Changes in CD4 count

Hypothesis testing

- Randomise into two groups

- Null hypothesis
  - No difference between treatments
  - Mean change in CD4 count is the same for A and B

- Alternative hypothesis
  - There is a difference between treatments

Hypothesis testing

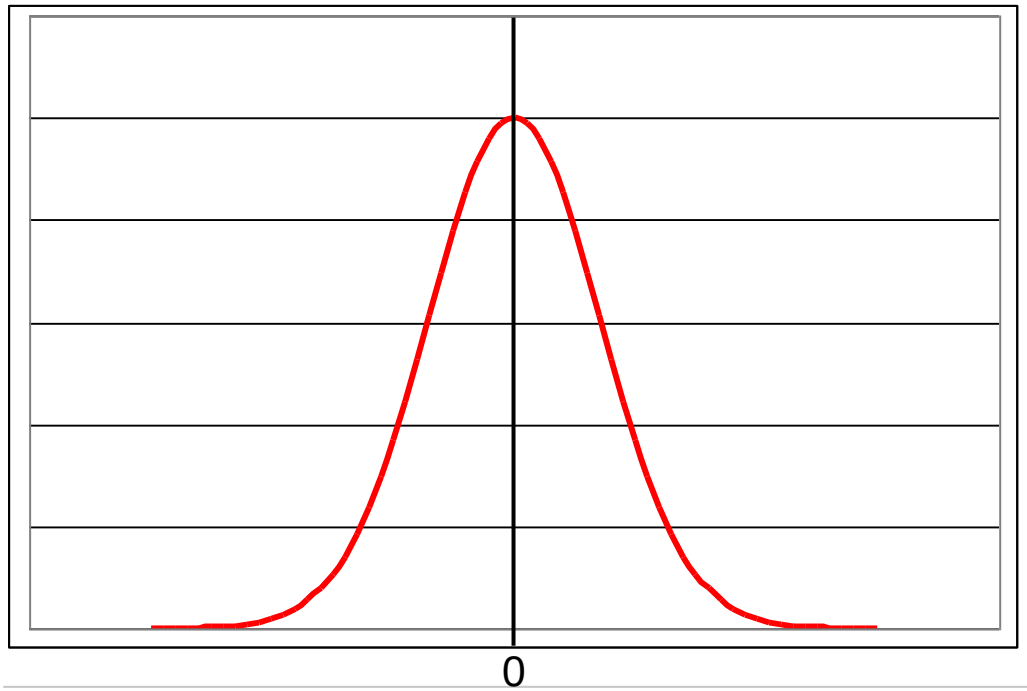
- Randomise into two groups

- Null hypothesis
  - No difference between treatments
  - Mean change in CD4 count is the same for A and B

- Alternative hypothesis
  - There is a difference between treatments

- Under the null hypothesis
  - The difference in mean change in CD4 count between A and B has a known probability distribution

Hypothesis testing



Mean difference in CD4: A-B

t-distribution

Hypothesis testing

- Randomise into two groups
- Null hypothesis
  - No difference between treatments
  - Mean change in CD4 count is the same for A and B
- Alternative hypothesis
  - There is a difference between treatments

- Under the null hypothesis
  - The difference in mean change in CD4 count between A and B has a known probability distribution
  - Calculate the probability of something as or more extreme than observed in our sample – p-value
  - If ‘p’ is small, we can reject the null hypothesis
  - If ‘p’ is not small, we can not reject the null hypothesis

Hypothesis testing

- Randomise into two groups
- Null hypothesis
    - No difference between treatments
    - Mean change in CD4 count is the same for A and B
- Alternative hypothesis
    - There is a difference between treatments
- Under the null hypothesis
    - The difference in mean change in CD4 count between A and B has a known probability distribution
    - Calculate the probability of something as or more extreme than observed in our sample – p-value
    - If ‘p’ is small, we can reject the null hypothesis
    - If ‘p’ is not small, we can not reject the null hypothesis

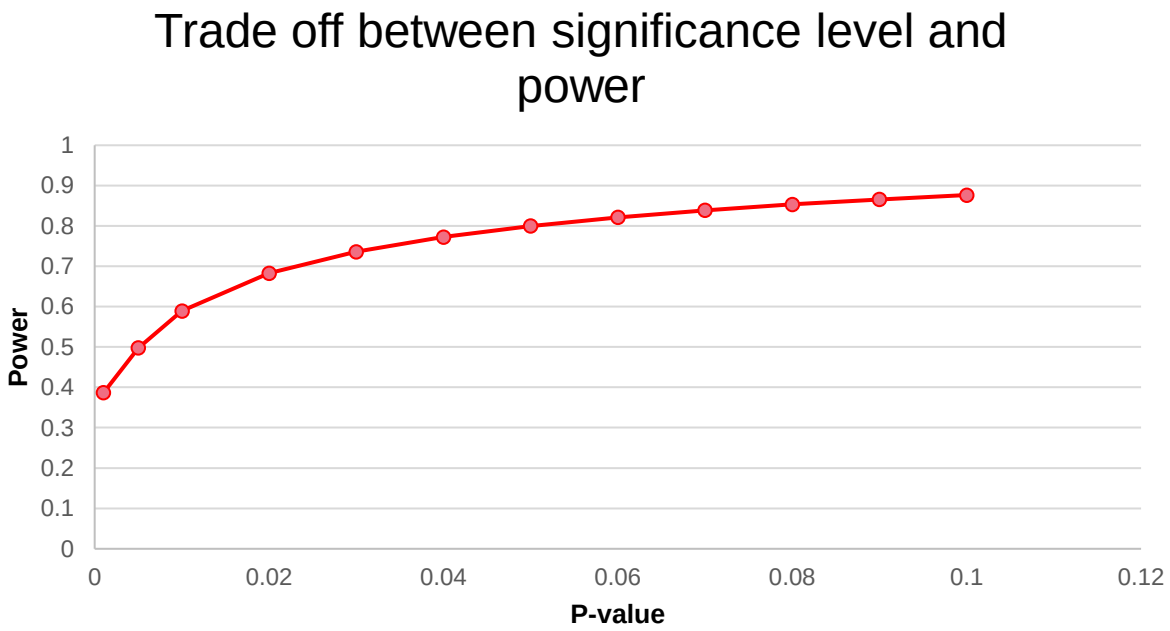
- **Important point**
    - **Failure to reject null hypothesis ≠ null hypothesis is true**

Hypothesis testing

- Type 1 error (size)
  - Reject the null hypothesis when it is true
  - 5%

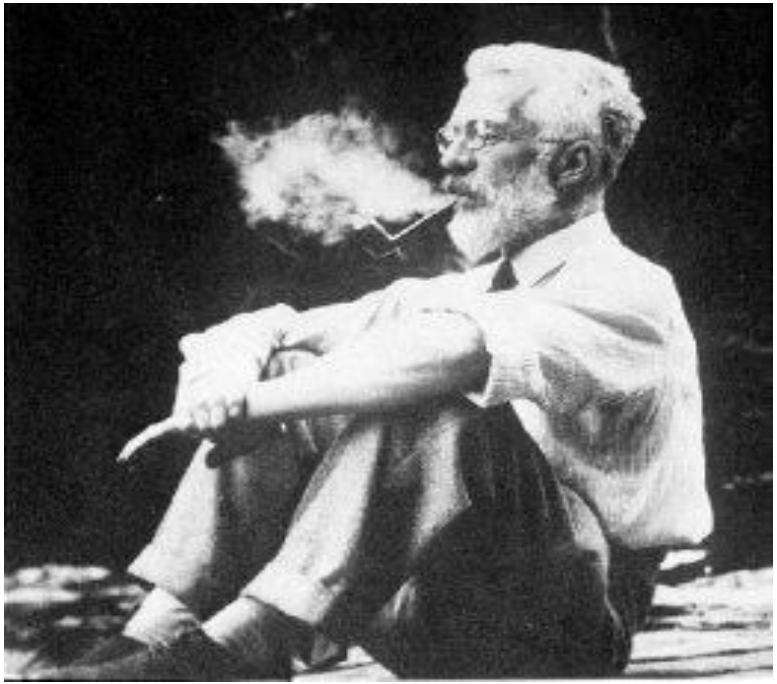
- Type 2 error
  - Fail to reject the null hypothesis when it is false
  - 1 - type 2 error = power

Hypothesis testing



Why 5%

- Ronald Fisher



Confidence intervals

- Estimate the difference between the treatments

- Calculate a range of values for the treatment difference which allows for random variation in your sample
    - A confidence interval

- The width of the confidence interval depends on the amount of random variation

Confidence intervals

- Formally not a probability statement
  - Probability treatment effect lies in a 95% CI ≠ 0.95
- If we repeated the trial 1,000 times, we'd expect the 95% CI to contain the treatment effect 950 times
  - 50 times (5%) won't – type 1 error
- Working interpretation
  - 95% CI gives a range of values for treatment effect that allows for random variation
  - NB Not bias

Good presentation of trial results

START trial

Table 2. Primary and Secondary End Points.*

End Point	Immediate-Initiation Group (N = 2326)		Deferred-Initiation Group (N = 2359)		Hazard Ratio (95% CI)†	P Value
	no.	*no./100 person-yr*	*no.*	*no./100 person-yr*		
Composite primary end point	42	0.60	96	1.38	0.43 (0.30–0.62)	<0.001

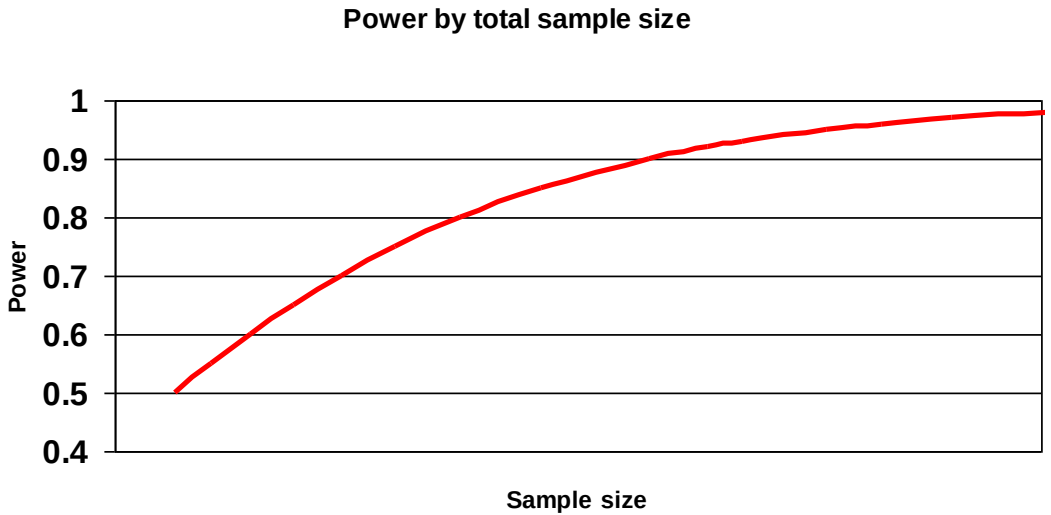
NEJM 2015;373:795-807

Contents

- Hypothesis testing & confidence intervals
- Power - small trials
- Randomisation and intention to treat
- Subgroup analyses
- Non-inferiority / equivalence trials

Sample size

- Power increases with larger sample size
- Turns out that power $\propto \sqrt{}$ total number of patients



Small trials

Very difficult to interpret

- **If they're negative, can't really interpret as no difference between treatments**
- **If results are positive, probably large overestimate**
    - (probably a Type 1 error)

Small trials

RCT comparing the effect of gemfibrozil and placebo on lowering triglycerides in HIV-positive people receiving antiretroviral treatments

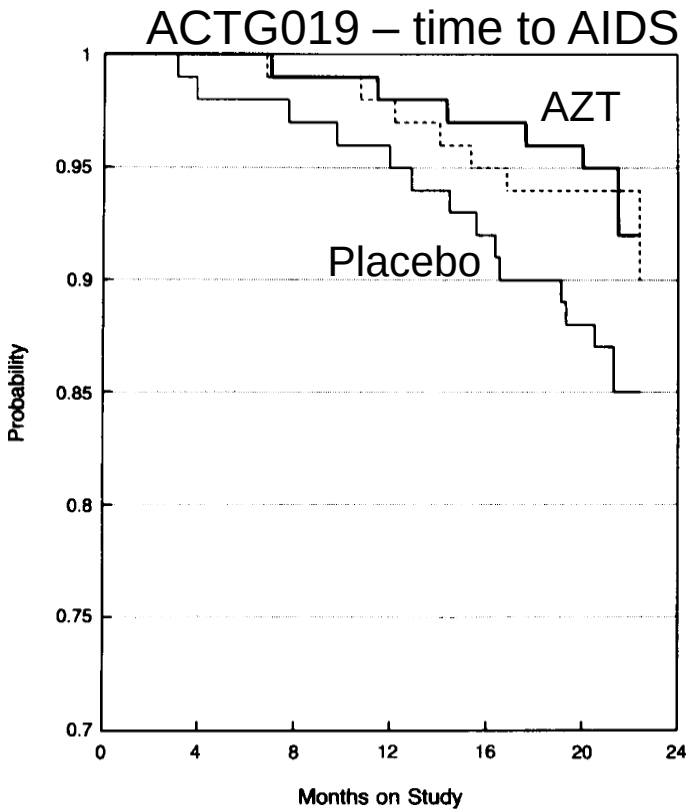
Table 2. Difference at week 16 between groups in mean change from baseline and week 4. Values are means ± SD.

Variable	Gemfibrozil group (n = 17)	Placebo group (n = 20)	Difference (95% CI)	*P*
Lipid				
Triglycerides (mmol/l)[a]				
Change from baseline	−0.88 (2.74)	0.12 (2.32)	−1.00 (−2.72 to 0.71)	0.24

Miller et al, *AIDS* 2002;16:2195-2200

Small trials

67% reduction in AIDS



Contents

- Hypothesis testing & confidence intervals
- Power - small trials
- Randomisation and intention to treat
- Subgroup analyses
- Non-inferiority / equivalence trials

Randomisation

Why?

- Being fair
- Being seen to be fair
- Basis of statistical inference

- Balances known and unknown confounders

Intention to treat

Randomise patients to two treatments

In analysis, compare patients according to their allocated treatment

- Ignore whether they refused, stopped or switched

Intention to treat

Justifications

- Answers the important question by comparing treatment policies
- Underestimates treatment effects, but by a small amount and in a known direction
- Retains randomisation - analyses by treatment received can be highly biased

Important implication

- Have to follow all randomised patients up
- No “withdrawals” – especially for stopping treatment

RCT of propanolol vs atenolol vs placebo in MI

Six week mortality rates

	Propanolol n=132	Atenolol n=127	Placebo n=129
Completed	(n=88)	(n=76)	(n=89)
Stopped	(n=44)	(n=51)	(n=40)
Total			

Wilcox et al, *BMJ* 1980

RCT of propanolol vs atenolol vs placebo in MI

Six week mortality rates

	Propanolol n=132	Atenolol n=127	Placebo n=129
Completed	3.4% (n=88)	2.6% (n=76)	11.2% (n=89)
Stopped			
Total			

Wilcox et al, *BMJ* 1980

RCT of propanolol vs atenolol vs placebo in MI

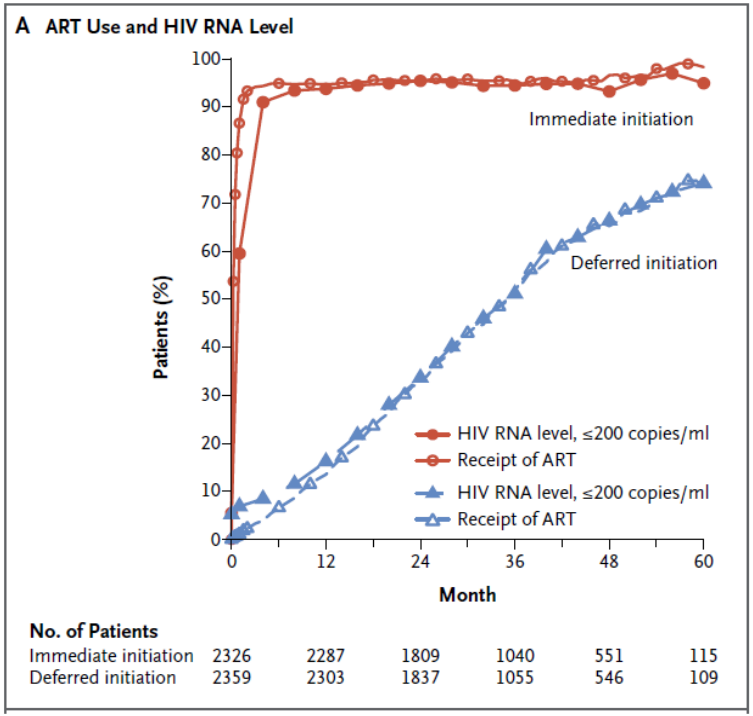
Six week mortality rates

	Propanolol n=132	**Atenolol** n=127	**Placebo** n=129
Completed	3.4% (n=88)	2.6% (n=76)	11.2% (n=89)
Stopped	15.9% (n=44)	17.6% (n=51)	12.5% (n=40)
Total	7.5%	8.6%	11.6%

Wilcox et al, *BMJ* 1980

Intention to treat

START trial



median CD4 at ART in deferred arm: 403 cells/mm^3

NEJM 2015;373:795-807

Contents

- Hypothesis testing & confidence intervals
- Power - small trials
- Randomisation and intention to treat
- Subgroup analyses
- Non-inferiority / equivalence trials

Subgroup analyses

Perform RCT comparing two treatments

Obtained overall results

Subgroup analyses

- Compare treatments in subgroups of all patients
- Goal to identify subgroups for whom the treatment is either most effective, or doesn’t work

Subgroup analyses

Problems

- Multiple treatment comparisons
  - Increased Type-I error
- Smaller sample sizes
  - Increased Type-II error

Worst possible combination, makes interpretation very difficult

Always some rational-sounding explanation after the fact

Subgroup analyses

ISIS-1 – aspirin vs placebo in acute MI (n>16,000)

Analyses by astrological birth sign

	% reduction odds death	p-value
Overall	15% (+/- 7%)	0.05

Subgroup analyses

ISIS-1 – aspirin vs placebo in acute MI (n>16,000)

Analyses by astrological birth sign

	% reduction odds death	p-value
Scorpio	48% (+/- 23%)	0.04
All others	12% (+/- 8%)	0.15
Overall	15% (+/- 7%)	0.05

Rgp120 Vaccine Study Group

Table 3. Attack rates of HIV-1 infection and vaccine efficacy (VE) against HIV-1 infection.

Category, parameter	Rate of HIV-1 infection		VE (95% CI)	*P*	
	Vaccine	Placebo		Unadjusted[a]	Adjusted[b]
All volunteers	241/3598 (6.7)	127/1805 (7.0)	6 (−17 to 24)	.59	>.5
Men	239/3391 (7.0)	123/1704 (7.2)	4 (−20 to 23)	.73	>.5
Women	2/207 (1.0)	4/101 (4.0)	74 (−42 to 95)	.093	.41
Race					
White (non-Hispanic)	211/2994 (7.0)	98/1495 (6.6)	−6 (−35 to 16)	.60	>.5
Men	211/2930 (7.2)	98/1468 (6.7)	−6 (−35 to 16)	.61	…
Women	0/64 (0)	0/27 (0)	…	…	…
Hispanic	14/239 (5.9)	9/128 (7.0)	15 (−96 to 63)	.70	>.5
Men	13/211 (6.2)	9/114 (7.9)	20 (−88 to 66)	.61	…
Women	1/28 (3.6)	0/14 (0)	…	…	…
Black (non-Hispanic)	6/233 (2.6)	9/116 (7.8)	67 (6 to 88)	.028	.24
Men	5/121 (4.1)	5/59 (8.5)	54 (−61 to 87)	.21	…
Women[c]	1/112 (0.9)	4/57 (7.0)	87 (−19 to 98)	.033	…
Asian (all men)	3/56 (5.4)	3/21 (14.3)	66 (−70 to 93)	.17	>.5
Other	7/76 (9.2)	8/45 (17.8)	50 (−39 to 82)	.18	>.5
Men	7/73 (9.6)	8/42 (19.0)	51 (−34 to 82)	.16	…
Nonwhite	30/604 (5.0)	29/310 (9.4)	47 (12 to 68)	.012	.13
Men	28/461 (6.1)	25/236 (10.6)	43 (3 to 67)	.036	…
Women	2/143 (1.4)	4/74 (5.4)	74 (−43 to 95)	.10	…
Age					
≤30 years	84/971 (8.7)	43/504 (8.5)	−1 (−46 to 30)	.95	>.5
>30 years	157/2627 (6.0)	84/1301 (6.5)	8 (−19 to 30)	.51	>.5
Education level[d]					
Less than a college degree	95/1409 (6.7)	52/713 (7.3)	8 (−29 to 34)	.63	>.5
College or graduate degree	146/2188 (6.7)	75/1092 (6.9)	4 (−27 to 27)	.77	>.5
Baseline behavioral risk score[e]					
Low risk	32/1211 (2.6)	11/609 (1.8)	−48 (−193 to 26)	.26	>.5
Medium risk	177/2229 (7.9)	90/1107 (8.1)	3 (−25 to 25)	.82	>.5
High risk	32/158 (20.3)	26/89 (29.2)	43 (4 to 66)	.032	.29

JID 2005;191:654-65

Rgp120 Vaccine Study Group

Table 3. Attack rates of HIV-1 infection and vaccine efficacy (VE) against HIV-1 infection.

Category, parameter	Rate of HIV-1 infection		VE (95% CI)	*P*	
	Vaccine	Placebo		Unadjusted[a]	Adjusted[b]
All volunteers	241/3598 (6.7)	127/1805 (7.0)	6 (−17 to 24)	.59	>.5
Men	239/3391 (7.0)	123/1704 (7.2)	4 (−20 to 23)	.73	>.5
Women	2/207 (1.0)	4/101 (4.0)	74 (−42 to 95)	.093	.41
Race					
White (non-Hispanic)	211/2994 (7.0)	98/1495 (6.6)	−6 (−35 to 16)	.60	>.5
Men	211/2930 (7.2)	98/1468 (6.7)	−6 (−35 to 16)	.61	…
Women	0/64 (0)	0/27 (0)	…	…	…
Hispanic	14/239 (5.9)	9/128 (7.0)	15 (−96 to 63)	.70	>.5
Men	13/211 (6.2)	9/114 (7.9)	20 (−88 to 66)	.61	…
Women	1/28 (3.6)	0/14 (0)	…	…	…
Black (non-Hispanic)	6/233 (2.6)	9/116 (7.8)	67 (6 to 88)	.028	.24
Men	5/121 (4.1)	5/59 (8.5)	54 (−61 to 87)	.21	…
Women[c]	1/112 (0.9)	4/57 (7.0)	87 (−19 to 98)	.033	…
Asian (all men)	3/56 (5.4)	3/21 (14.3)	66 (−70 to 93)	.17	>.5
Other	7/76 (9.2)	8/45 (17.8)	50 (−39 to 82)	.18	>.5
Men	7/73 (9.6)	8/42 (19.0)	51 (−34 to 82)	.16	…
Nonwhite	30/604 (5.0)	29/310 (9.4)	47 (12 to 68)	.012	.13
Men	28/461 (6.1)	25/236 (10.6)	43 (3 to 67)	.036	…
Women	2/143 (1.4)	4/74 (5.4)	74 (−43 to 95)	.10	…
Age					
≤30 years	84/971 (8.7)	43/504 (8.5)	−1 (−46 to 30)	.95	>.5
>30 years	157/2627 (6.0)	84/1301 (6.5)	8 (−19 to 30)	.51	>.5
Education level[d]					
Less than a college degree	95/1409 (6.7)	52/713 (7.3)	8 (−29 to 34)	.63	>.5
College or graduate degree	146/2188 (6.7)	75/1092 (6.9)	4 (−27 to 27)	.77	>.5
Baseline behavioral risk score[e]					
Low risk	32/1211 (2.6)	11/609 (1.8)	−48 (−193 to 26)	.26	>.5
Medium risk	177/2229 (7.9)	90/1107 (8.1)	3 (−25 to 25)	.82	>.5
High risk	32/158 (20.3)	26/89 (29.2)	43 (4 to 66)	.032	.29

JID 2005;191:654-65

Rgp120 Vaccine Study Group

Table 3. Attack rates of HIV-1 infection and vaccine efficacy (VE) against HIV-1 infection.

Category, parameter	Rate of HIV-1 infection		VE (95% CI)	*P*	
	Vaccine	Placebo		Unadjusted[a]	Adjusted[b]
All volunteers	241/3598 (6.7)	127/1805 (7.0)	6 (−17 to 24)	.59	>.5
Men	239/3391 (7.0)	123/1704 (7.2)	4 (−20 to 23)	.73	>.5
Women	2/207 (1.0)	4/101 (4.0)	74 (−42 to 95)	.093	.41
Race					
White (non-Hispanic)	211/2994 (7.0)	98/1495 (6.6)	−6 (−35 to 16)	.60	>.5
Men	211/2930 (7.2)	98/1468 (6.7)	−6 (−35 to 16)	.61	…
Women	0/64 (0)	0/27 (0)	…	…	…
Hispanic	14/239 (5.9)	9/128 (7.0)	15 (−96 to 63)	.70	>.5
Men	13/211 (6.2)	9/114 (7.9)	20 (−88 to 66)	.61	…
Women	1/28 (3.6)	0/14 (0)	…	…	…
Black (non-Hispanic)	6/233 (2.6)	9/116 (7.8)	67 (6 to 88)	.028	.24
Men	5/121 (4.1)	5/59 (8.5)	54 (−61 to 87)	.21	…
Women[c]	1/112 (0.9)	4/57 (7.0)	87 (−19 to 98)	.033	…
Asian (all men)	3/56 (5.4)	3/21 (14.3)	66 (−70 to 93)	.17	>.5
Other	7/76 (9.2)	8/45 (17.8)	50 (−39 to 82)	.18	>.5
Men	7/73 (9.6)	8/42 (19.0)	51 (−34 to 82)	.16	…
Nonwhite	30/604 (5.0)	29/310 (9.4)	47 (12 to 68)	.012	.13
Men	28/461 (6.1)	25/236 (10.6)	43 (3 to 67)	.036	…
Women	2/143 (1.4)	4/74 (5.4)	74 (−43 to 95)	.10	…
Age					
≤30 years	84/971 (8.7)	43/504 (8.5)	−1 (−46 to 30)	.95	>.5
>30 years	157/2627 (6.0)	84/1301 (6.5)	8 (−19 to 30)	.51	>.5
Education level[d]					
Less than a college degree	95/1409 (6.7)	52/713 (7.3)	8 (−29 to 34)	.63	>.5
College or graduate degree	146/2188 (6.7)	75/1092 (6.9)	4 (−27 to 27)	.77	>.5
Baseline behavioral risk score[e]					
Low risk	32/1211 (2.6)	11/609 (1.8)	−48 (−193 to 26)	.26	>.5
Medium risk	177/2229 (7.9)	90/1107 (8.1)	3 (−25 to 25)	.82	>.5
High risk	32/158 (20.3)	26/89 (29.2)	43 (4 to 66)	.032	.29

“The efficacy trends in subgroups may provide clues for the development of effective immunization approaches”.

JID 2005;191:654-65

Good presentation - START

Subgroup	Percentage in Group	Immediate Initiation	Deferred Initiation	Hazard Ratio (95% CI)	P Value for Interaction
		no. of patients with event (rate per 100 person-yr)			
Age					0.98
≤35 yr	48.8	15 (0.43)	31 (0.91)	0.47	
>35 yr	51.2	27 (0.78)	65 (1.85)	0.42	
Sex					0.38
Male	73.2	35 (0.66)	74 (1.40)	0.47	
Female	26.8	7 (0.42)	22 (1.34)	0.31	
Baseline HIV RNA					0.25
<5000 copies/ml	31.8	12 (0.56)	18 (0.83)	0.66	
5000–30,000 copies/ml	35.5	13 (0.53)	36 (1.41)	0.38	
>30,000 copies/ml	32.5	17 (0.72)	42 (1.92)	0.37	
Smoker					0.93
Yes	31.9	18 (0.78)	43 (1.81)	0.43	
No	68.1	24 (0.52)	53 (1.16)	0.44	
Framingham 10-yr CHD risk					0.56
<0.8	32.7	8 (0.35)	17 (0.77)	0.46	
0.8–3.6	32.3	11 (0.48)	27 (1.23)	0.39	
>3.6	33.5	23 (1.00)	50 (2.05)	0.50	

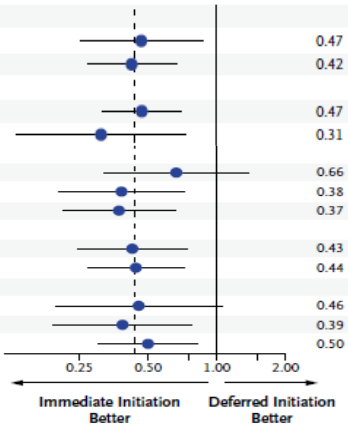


Figure 3. Subgroup Analyses for the Primary End Point.
For subgroups that were defined according to age, CD4+ count, HIV RNA level, and risk of coronary heart disease (CHD), the continuous variables were used for interaction tests. For 71 patients (1.5%), the Framingham Heart Study risk of CHD could not be calculated because of missing data. Of the patients with missing data, the primary end point occurred in 2 in the deferred-initiation group.

NEJM 2015;373:795-807

Contents

- Hypothesis testing & confidence intervals
- Power - small trials
- Randomisation and intention to treat
- Subgroup analyses
- Non-inferiority / equivalence trials

Hypothetical example

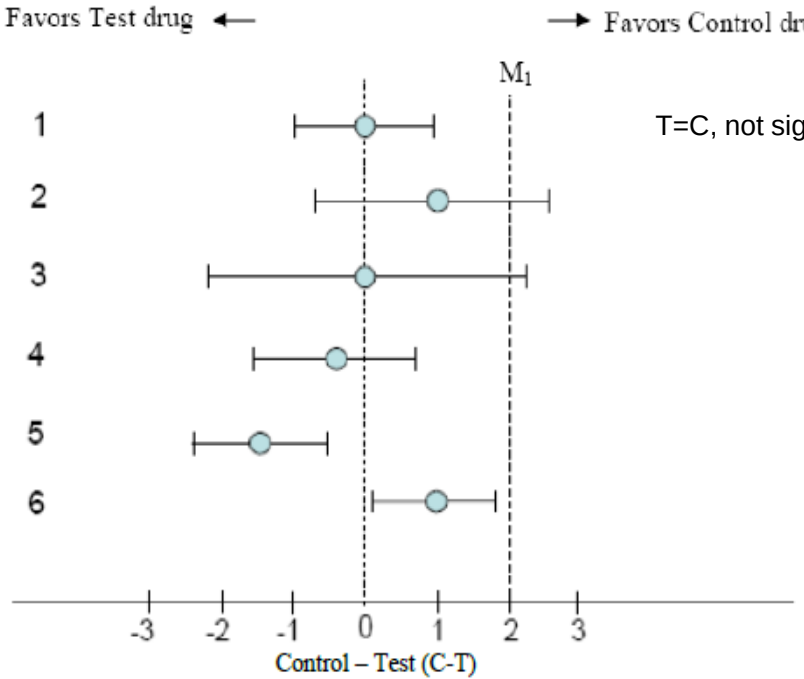
- Results:
  - A : 40/50 (80%) patients undetectable HIV (<200c/ml)
  - B : 39/50 (78%) patients undetectable HIV
- Difference between arms
  - -2%, 95% CI **-18% to 14%**, p=0.806
- Is this sufficient evidence that treatment B is as good as A?
  - Or at least no worse?

Framework for non-inferiority trials

- Not a hypothesis testing approach
  - Statistical significance not important

- To conclude non-inferiority
  - Need to shrink the lower 95% confidence limit on the treatment difference within some small amount (that everyone agrees on)
  - Non-inferiority delta

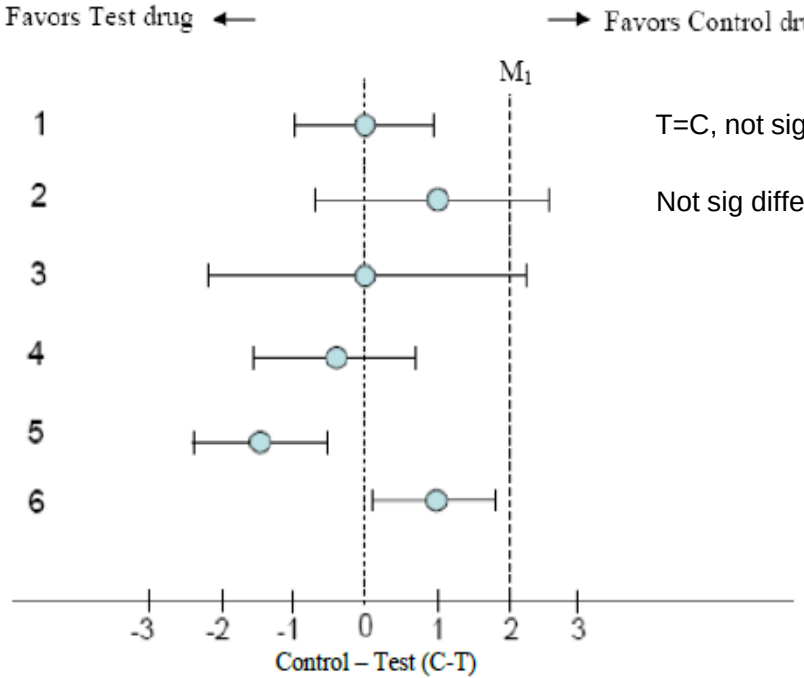
Interpreting different possible trial outcomes



T=C, not sig different, T non-inferior

https://www.fda.gov/downloads/Drugs/Guidances/UCM202140.pdf

Interpreting different possible trial outcomes

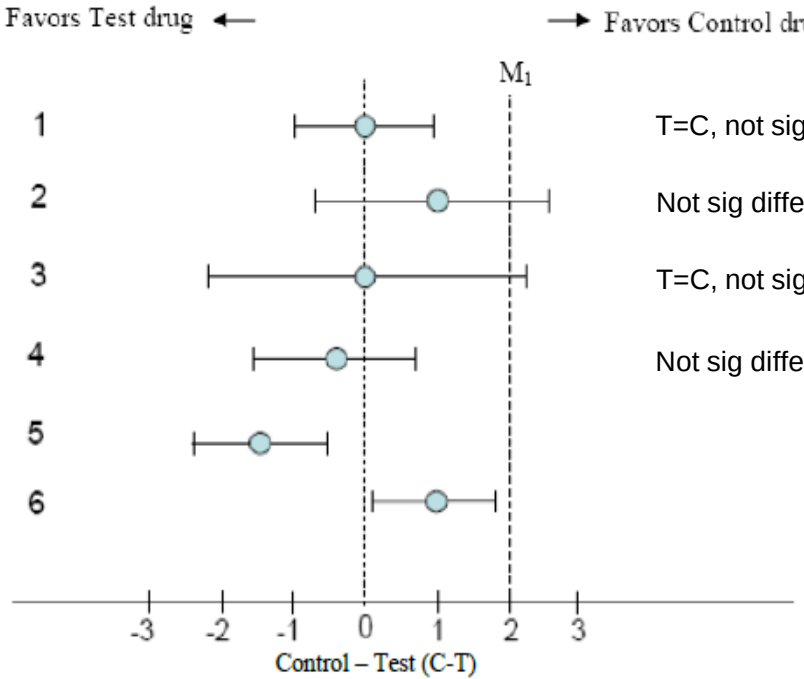


T=C, not sig different, T non-inferior

Not sig different, but non-inferiority not proven

https://www.fda.gov/downloads/Drugs/Guidances/UCM202140.pdf

Interpreting different possible trial outcomes



T=C, not sig different, T non-inferior

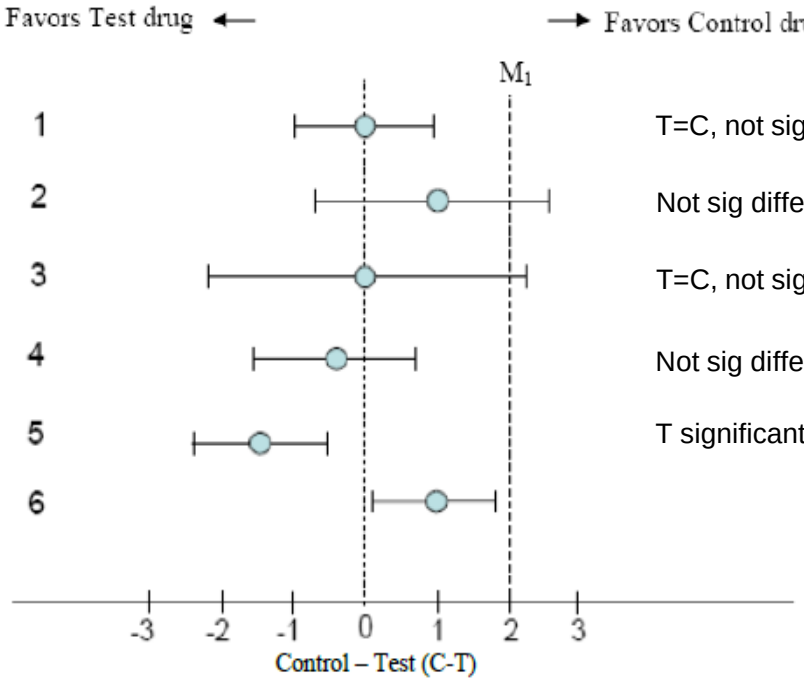
Not sig different, but non-inferiority not proven

T=C, not sig different, but non-inferiority not proven

Not sig different, and T non-inferior

https://www.fda.gov/downloads/Drugs/Guidances/UCM202140.pdf

Interpreting different possible trial outcomes



T=C, not sig different, T non-inferior

Not sig different, but non-inferiority not proven

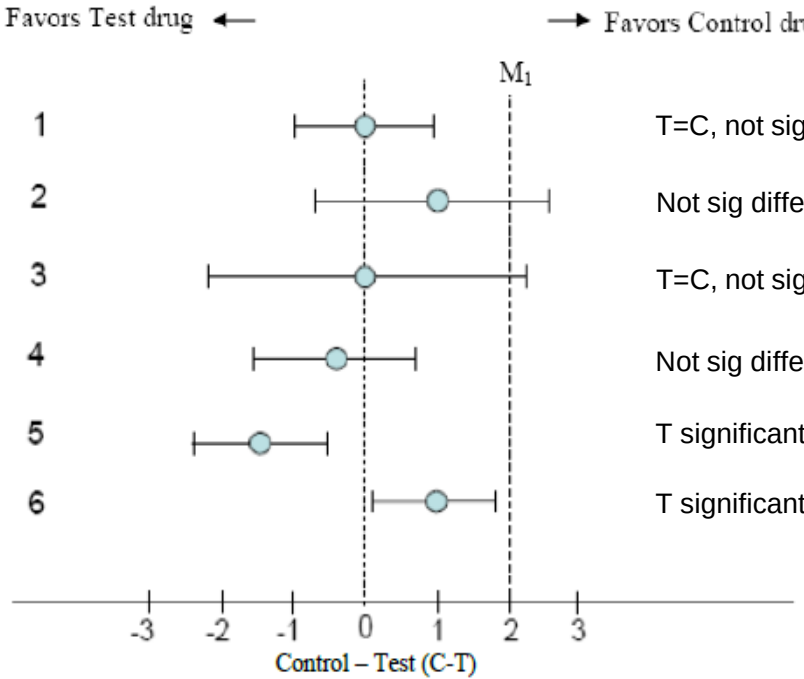
T=C, not sig different, but non-inferiority not proven

Not sig different, and T non-inferior

T significantly better than C

https://www.fda.gov/downloads/Drugs/Guidances/UCM202140.pdf

Interpreting different possible trial outcomes



T=C, not sig different, T non-inferior

Not sig different, but non-inferiority not proven

T=C, not sig different, but non-inferiority not proven

Not sig different, and T non-inferior

T significantly better than C

T significantly worse than C, but non-inferior

https://www.fda.gov/downloads/Drugs/Guidances/UCM202140.pdf

Second-Line trial

- Compared LPVr+2NRTIs (SOC) vs RTG+LPVr

- Primary endpoint
  - Undetectable viral load (<200 copies/mL) at week 48

- Wanted to establish RTG+LPVr was non-inferior (no worse) than SOC
  - Sample size based on a non-inferiority delta of 12%
  - Expected 80% undetectable viral load in both arms

Lancet 2013; 381: 2091-99

Second-Line trial

- 271 participants randomized to SOC arm and 270 to RTG arm

	Control group	Raltegravir group	Difference (95% CI)
Plasma viral load <200 copies/mL			
Modified intention-to-treat (primary outcome)	219/271 (80·8%)	223/270 (82·6%)	1·8 (–4·7 to 8·3)
Baseline viral load ≤100 000 copies per mL*	188/219 (85·8%)	184/210 (87·6%)	1·8 (–4·6 to 8·2)
Baseline viral load >100 000 copies per mL*	31/52 (59·6%)	39/60 (65·0%)	5·4 (–12·6 to 23·4)
Per protocol	211/249 (84·7%)	211/246 (85·8%)	1·0 (–5·2 to 7·3)
Non-completer classed as failure†	208/271 (76·8%)	210/270 (77·8%)	1·0 (–6·0 to 8·1)
Plasma viral load <50 copies per mL			
Modified intention to treat	191/271 (70·5%)	192/270 (71·1%)	0·6 (–7·0 to 8·3)
Per protocol	183/249 (73·5%)	185/246 (75·2%)	1·7 (–6·0 to 9·4)
Non-completer classed as failure†	180/271 (66·4%)	184/270 (68·1%)	1·7 (–6·2 to 9·6)

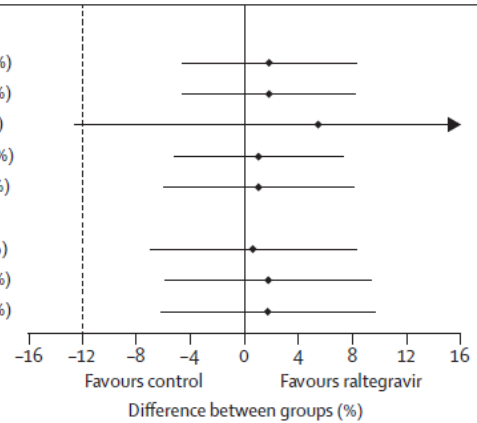


Figure 3: **Virological response at week 48, stratified by baseline viral load and analytical population**
The non-inferiority margin is –12. *Based on samples tested locally. †Equivalent to the FDA snapshot analysis

Lancet 2013; 381: 2091-99

Final comment

- RCTs are extremely powerful

- Have two well defined treatments (that are reasonably different)
- Randomise lots of subjects
- Follow-them all up

- And you will get the right answer

Lancet 2013; 381: 2091-99

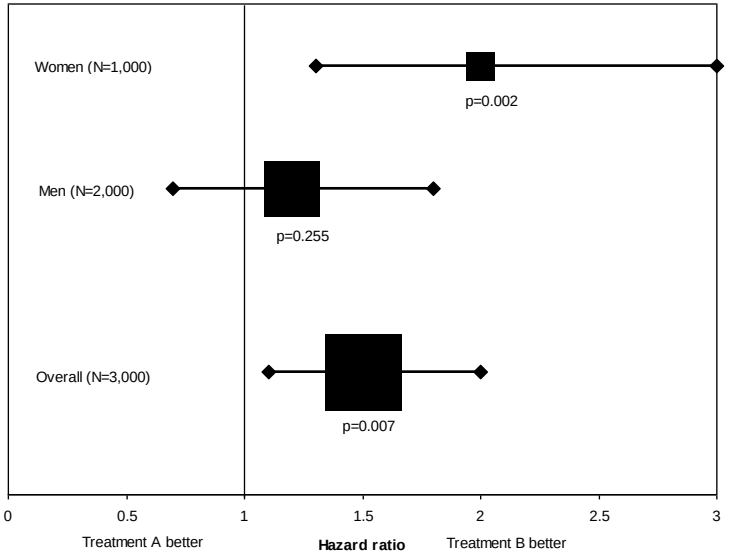
Thank you

Participant question

Supplementary

An RCT presents results summarised in the figure

Figure 1. Comparison of treatment A and B on death rates, overall and by sex



What is the best interpretation of these results?

1. Treatment B works better in women
2. Treatment B works in women but not in men
3. The estimated treatment effect in men and women is consistent