Proportions in medical research

Stephen Kerr

Summary

- How do we commonly use proportions in medical research?

- Relative risk
- Risk difference
- Odds ratio
- NNT/NNH

- Chi-square and Fisher’s exact test – Formal comparisons

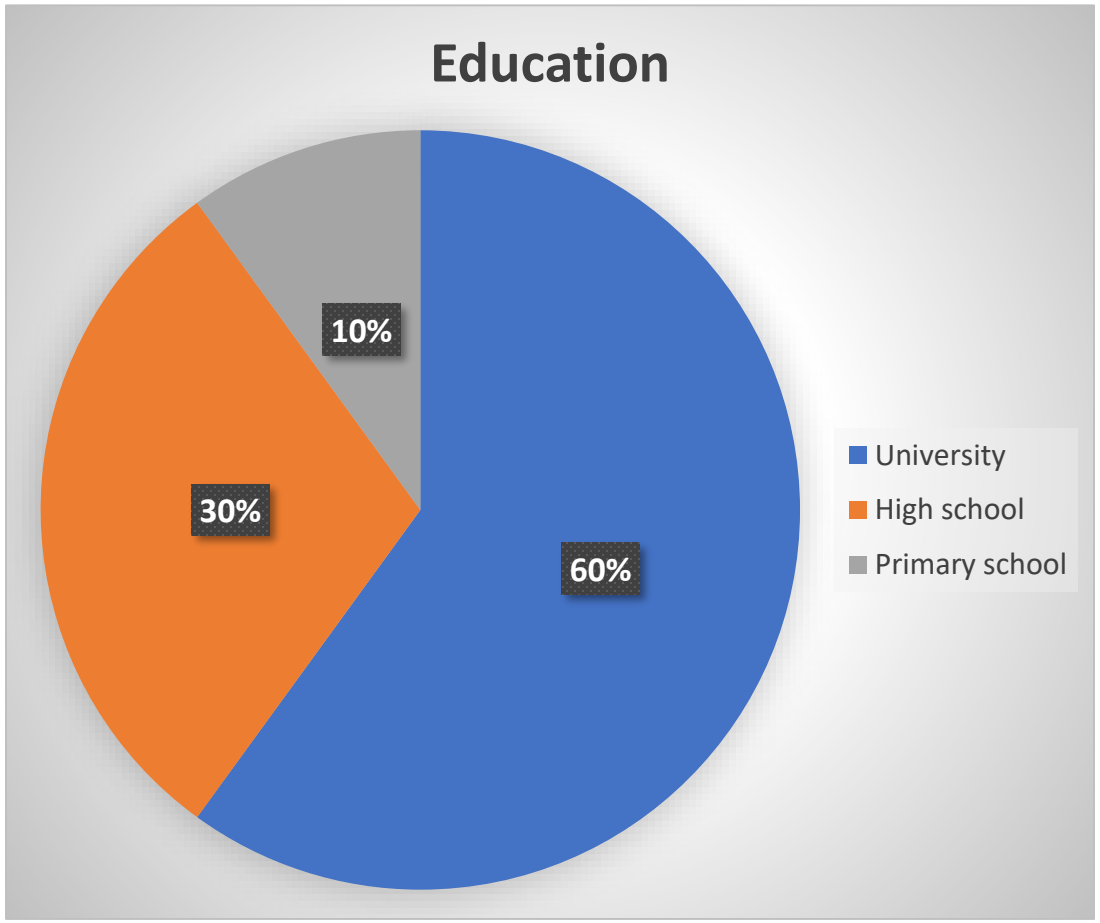
- Multivariate regression models to compare proportions

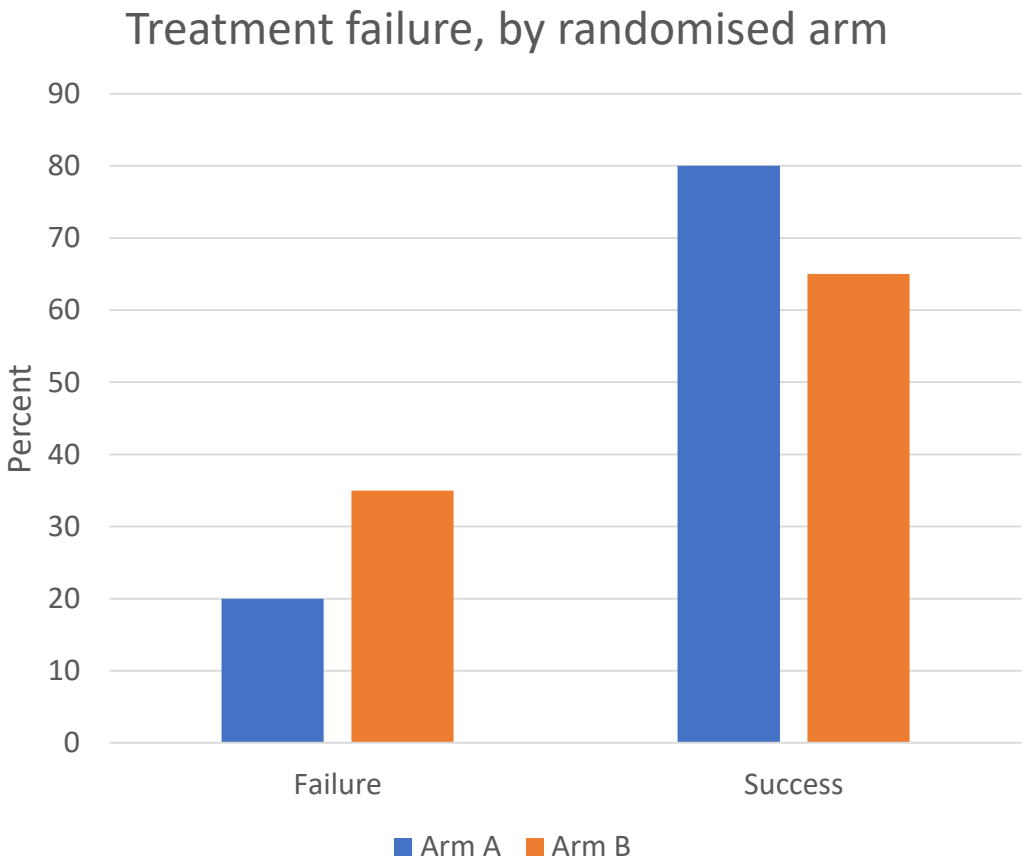
Proportions

- A fraction or part of a whole
- Can expressed as proportion (0 - 1) or percent (0-100%)

- Commonly used in medical science
- N (%) of categories in describing results
- If there’s 2 categories - for summarizing results (eg relative risk, absolute risk difference, NNT/NNH, odds ratio)
- Logistic regression (for multivariate risk)
- Multinomial or ordinal logistic regression

Proportions by group





Risk, and no effect values

- For a risk difference - Risk in group A – risk in group B

  eg 5% - 5% = 0% ; 0 is the no effect value

  So, if the 95%CI around the risk difference doesn't cross 0, result is significant at $P<0.05$.

- For a relative risk – risk in group A / risk in group B

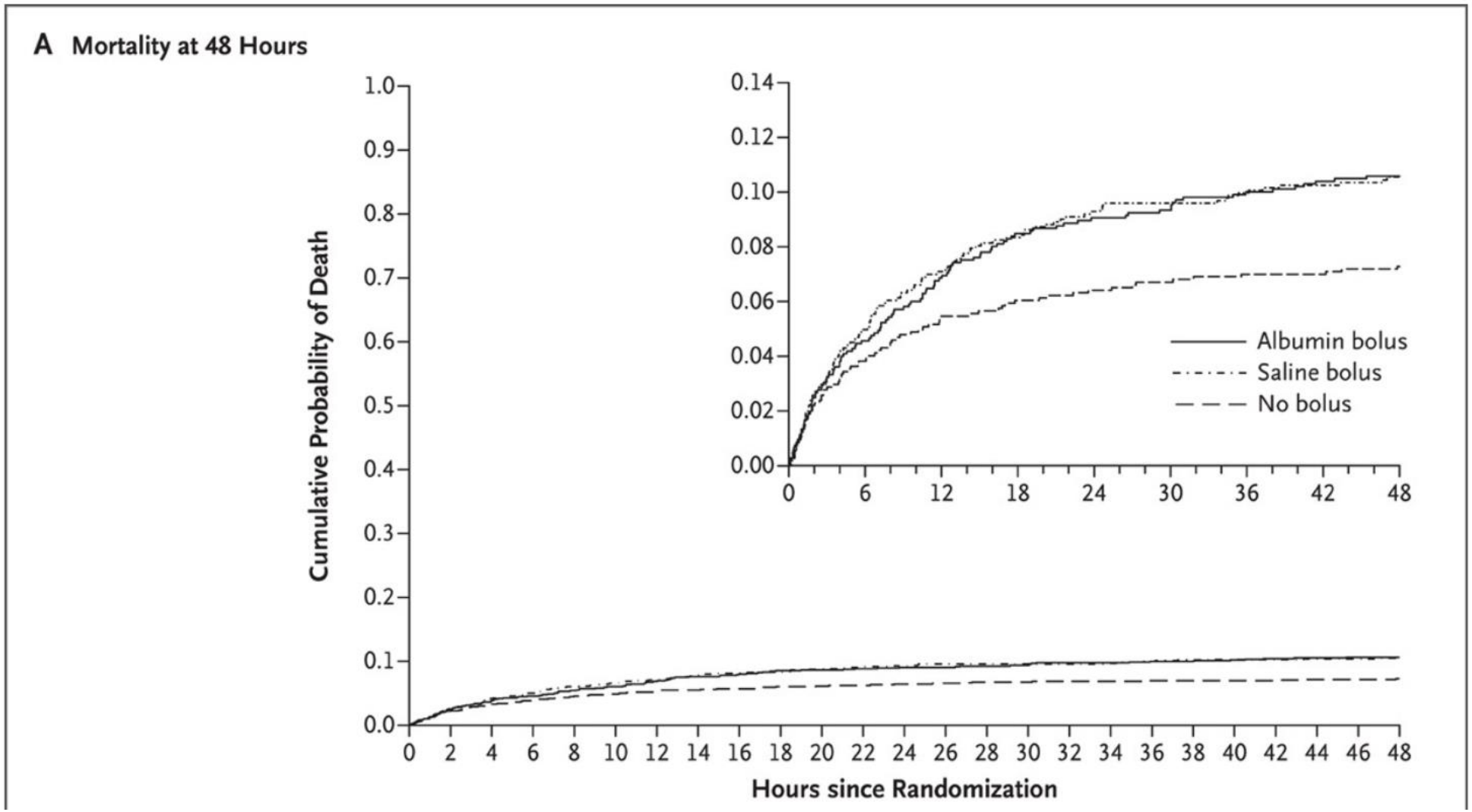
  eg 5%/5% = 1; 1 is the no effect value

  So, if the 95%CI around the risk difference doesn't cross 1, result is significant at $P<0.05$.

ORIGINAL ARTICLE

Mortality after Fluid Bolus in African Children with Severe Infection

Kathryn Maitland, M.B., B.S., Ph.D., Sarah Kiguli, M.B., Ch.B., M.Med., Robert O. Opoka, M.B., Ch.B., M.Med., Charles Engoru, M.B., Ch.B., M.Med., Peter Olupot-Olupot,



Risk difference (absolute risk)

Arm	N (%) who died	N (%) who survived	Total
Albumin bolus	111 (10.6)	939 (89.4)	1050
Saline bolus	110 (10.5)	937 (89.5)	1047
No bolus	76 (7.3)	968 (92.7)	1044

- Risk difference (saline vs no bolus)

  = 10.5 – 7.3 = 3.2% (95%CI 0.08 – 5.7%); P = 0.01

- For every 100 children receiving a saline bolus, we would expect 3.2 additional deaths at 48 hours, compared to the number we would have expected had none received a bolus

Maitland et al Mortality after Fluid Bolus in African Children with Severe Infection. NEJM (2011) 364:2483-95

Slide from Caroline Sabin, UCL

Relative risk (risk ratio)

Arm	N (%) who died	N (%) who survived	Total
Albumin bolus	111 (10.6)	939 (89.4)	1050
Saline bolus	110 (10.5)	937 (89.5)	1047
No bolus	76 (7.3)	968 (92.7)	1044

- Relative risk (saline vs no bolus)

  = 10.5 ÷ 7.3 = 1.44% (95%CI 1.09 – 1.91%); P = 0.01

- Risk of dying in the children who received the saline bolus is 44% higher than the risk in the no bolus group.
- OR – risk of dying in the saline bolus group is 1.44 x risk in the no bolus group

Slide from Caroline Sabin, UCL

Odds ratio (OR)

Arm	N (%) who died	N (%) who survived	Total
Saline bolus	110 (10.5)	937 (89.5)	1047
No bolus	76 (7.3)	968 (92.7)	1044

- RR – compares the proportion of endpoints between groups
- Odds – number of endpoints for everyone who doesn't have the endpoint
- In case-control studies and cross-sectional studies (logistic regression)
  - comparisons between groups made with OR
  - Not really intuitive for most people

- **OR (saline vs no bolus)**
- **= 110/937 ÷ 76/968 = 1.5 (96%CI 1.08 – 2.06); P = 0.01**

Slide from Caroline Sabin, UCL

Risk difference vs RR & OR

RD = 3.2 (95%CI 0.08 – 5.7)%; P = 0.01

RR = 1.44 (95%CI 1.09 – 1.91); P = 0.01

OR = 1.5 (95%CI CI 1.08 – 2.06); P= 0.01

VERY different information conveyed

- RR – how much more or less likely is something to occur in one group vs another group (strength of association)

- RD – what is the difference in the outcomes between the groups ***or how much will your intervention or risk factor change the outcome***

Relative risk

- RR = 0.5%/0.1% = 5.0
- RR = 50%/10% = 5.0

Versus risk difference

- But it doesn’t tell us anything about the background rate on which the comparison is based.

- RD = 0.5% – 0.1% = 0.4% (not much)
- RD = 50% – 10% = 40% (a lot!)

British Doctors Study

Disease	Smoking Status	Mortality rate/10^5py	RR	AR/10^5py
Lung Cancer	Yes	140	= 140/10 = 14	= 140-10 = 130
	No	10		
Coronary heart disease	Yes	669	= 669/413 = 1.6	= 669-413 = 260
	No	413		

BBC | Sign in | News | Sport

NEWS

Home | Video | World | Asia | UK | Business | T

It’s report said 50g of processed meat a day – less than 2 slices of bacon – increased the chance of developing colorectal cancer by 18%

Health

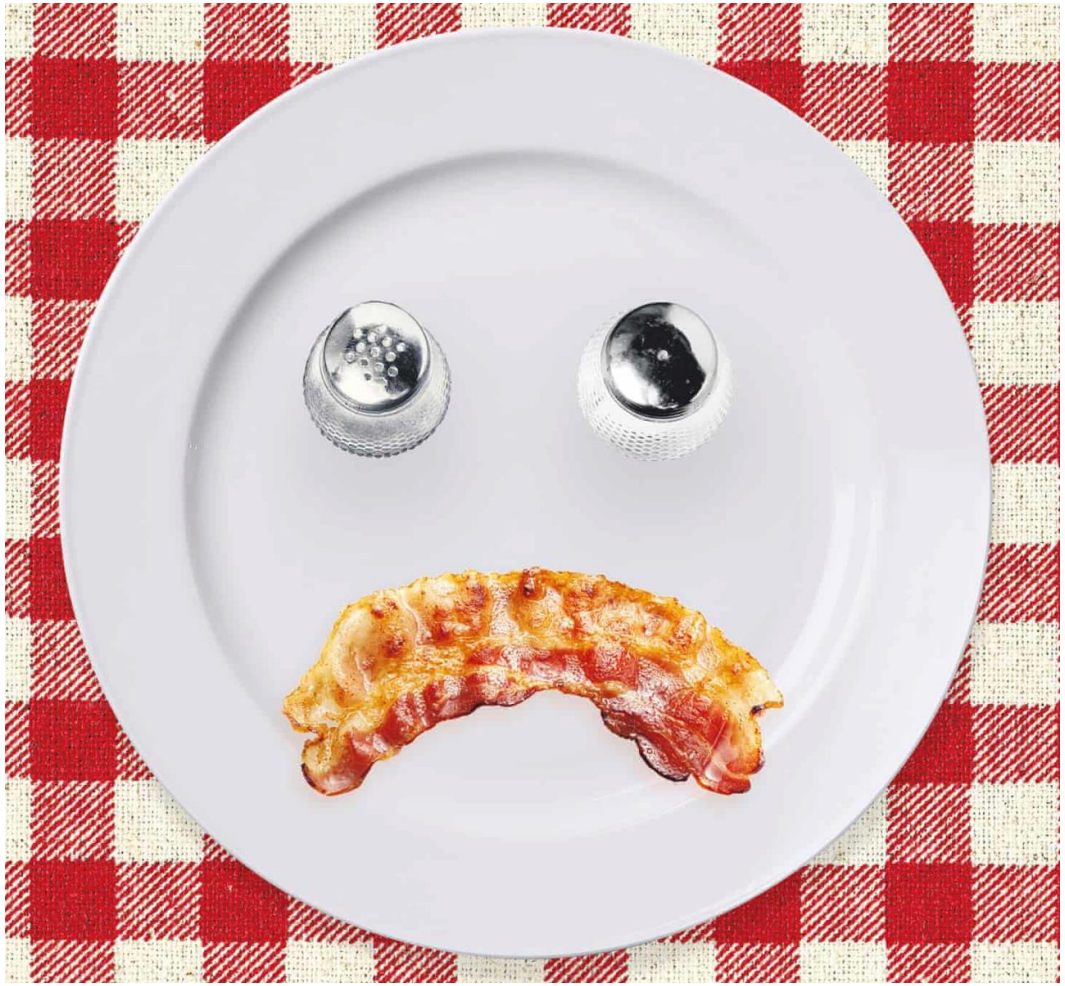
Processed meats do cause cancer - WHO

By James Gallagher
Health editor, BBC News website

26 October 2015 | Share

Processed meats - such as bacon, sausages and ham - do cause cancer, according to the World Health Organization (WHO).

Its report said 50g of processed meat a day - less than two slices of bacon - increased the chance of developing colorectal cancer by 18%.



The long read

Yes, bacon really is killing us

Decades' worth of research proves that chemicals used to make bacon do cause cancer. So how did the meat industry convince us it was safe? By Bee Wilson

There was a little cafe I used to go to that did the best bacon sandwiches. They came in a soft and pillowy white bap. The bacon, thick-cut from a local butcher, was midway between crispy and chewy. Ketchup and HP sauce were served in miniature jars with the sandwich, so you could dab on the exact amount you liked. That was all there was to it: just bread and bacon and sauce. Eating one of these sandwiches, as I did every few weeks, with a cup of strong coffee, felt like an uncomplicated pleasure.

And then, all of a sudden, the bacon sandwich stopped being quite so comforting. For a few weeks in October 2015, half the people I knew were talking about the news that eating bacon was now a proven cause of cancer. You couldn't miss the story: it was splashed large in every newspaper and all over the web. As one journalist wrote in Wired, "Perhaps no two words together are more likely to set the internet aflame than BACON and CANCER." The BBC website announced, matter-of-factly, that "Processed meats do cause c[illegible] the Sun went with "Banger out of Order" and "Killer in the Kitchen".

Screenshot

In the weeks following news of the WHO report, sales of bacon and sausages fell dramatically. British supermarkets reported a £3m drop in sales in just a fortnight. ("It was very detrimental," said Kirsty Adams, the product developer for meat at Marks and Spencer.)

Example: processed meat and bowel cancer

What does a 18% increased risk of bowel cancer really mean?

= 20 people

for people who eat the least processed meat...

...56 of 1000 will develop bowel cancer

absolute risk is 5.6%

+ 18% relative risk

+ 1% absolute risk

for people who eat the most processed meat...

...66 of 1000 will develop bowel cancer

absolute risk increases to 6.6%

VISUALIZING HEALTH

A Scientifically Vetted Style Guide for Communicating Health Data

About | The Wizard | Browse the Gallery | Using Visualizing Health

What & Why

Data can help people make better decisions about their health. But it's not always clear what sort of presentation will make the most sense to the most people. ... **READ MORE**

How

Visualizing Health contains graphic displays of health information that we've validated through research among the general public. ... **READ MORE**

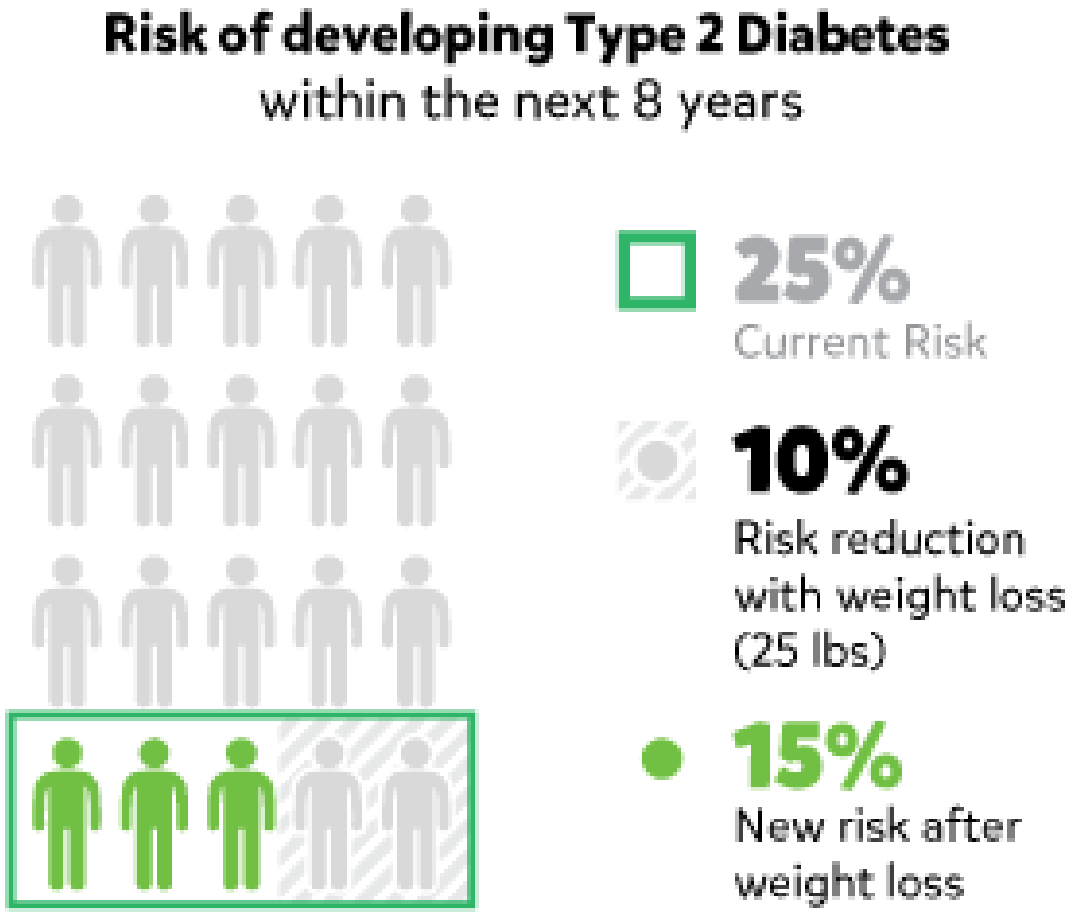
Who

Visualizing Health was developed by the University of Michigan, with funding from the Robert Wood Johnson Foundation. ... **READ MORE**

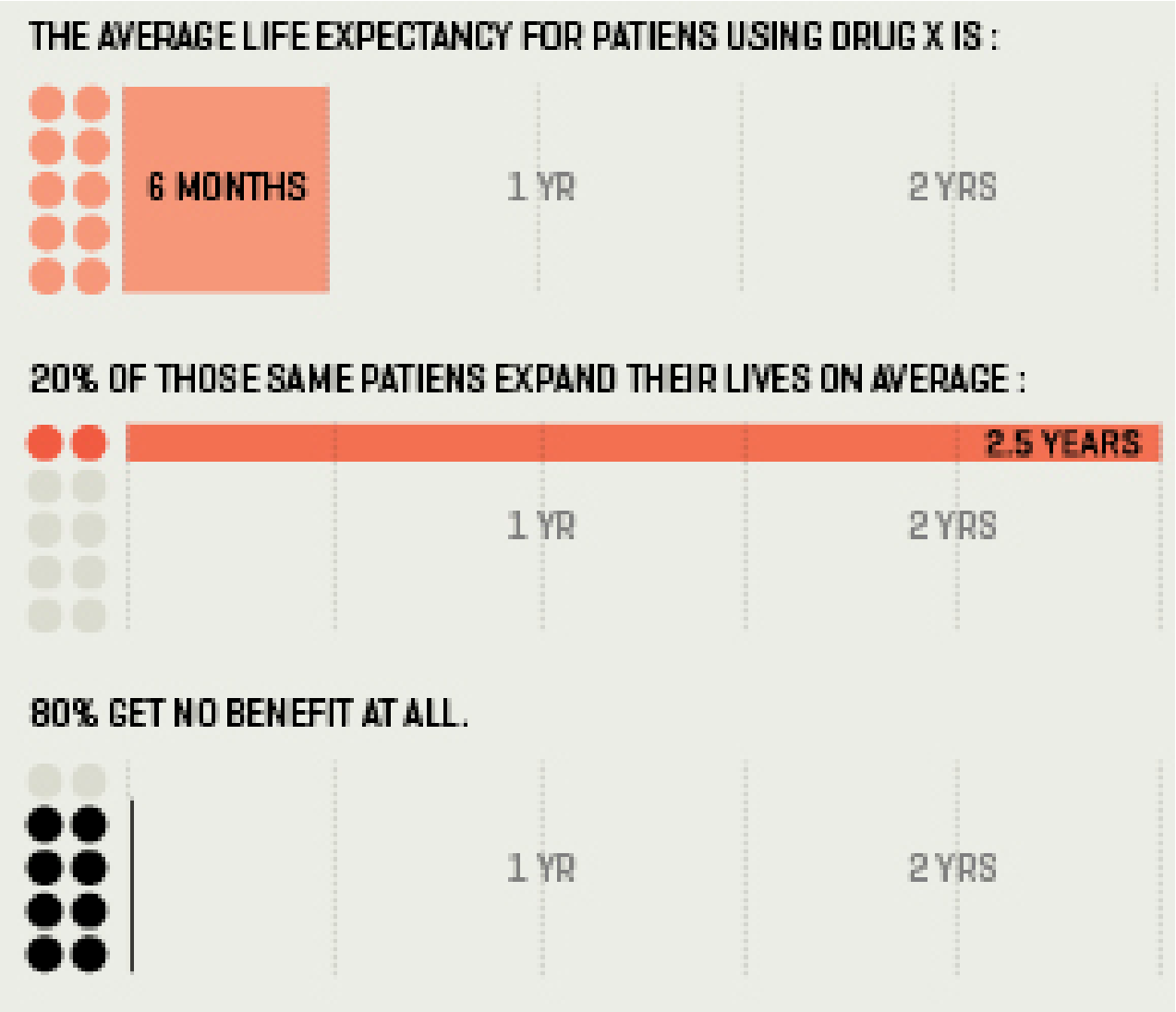
UNIVERSITY OF MICHIGAN

Support provided by Robert Wood Johnson Foundation

Benefits of risk reduction



Years of life saved by taking a drug



Number needed to treat (NNT)
Number needed to harm (NNH)

- Number of people who need to be treated with the intervention over period of study to prevent/cause one adverse event

- NNT = 100/RD (%) = 100/3.2 = 31.25

- For every 32 children who receive a saline bolus, there will be one additional death within 48 hours

How do we get a p-value for RR, RD and OR?

- Chi-square test
- Fisher's exact test
- The null hypothesis is that the categorical variables are independent (their frequency distribution is the same at all levels of the variable)

Key
frequency
expected frequency
row percentage

Expected cell frequency must be ≥5 for a chi-square test

row	col: Dead	Alive	Total
Saline Bolus	110	937	1,047
	93.1	953.9	1,047.0
	10.51	89.49	100.00
No bolus	76	968	1,044
	92.9	951.1	1,044.0
	7.28	92.72	100.00
Total	186	1,905	2,091
	186.0	1,905.0	2,091.0
	8.90	91.10	100.00

Expected count = Row total x column total/total number

= (1047x186)/2091 = 93.1

```
Pearson chi2(1) =   6.7152   Pr = 0.010
 Fisher's exact =                0.011
```

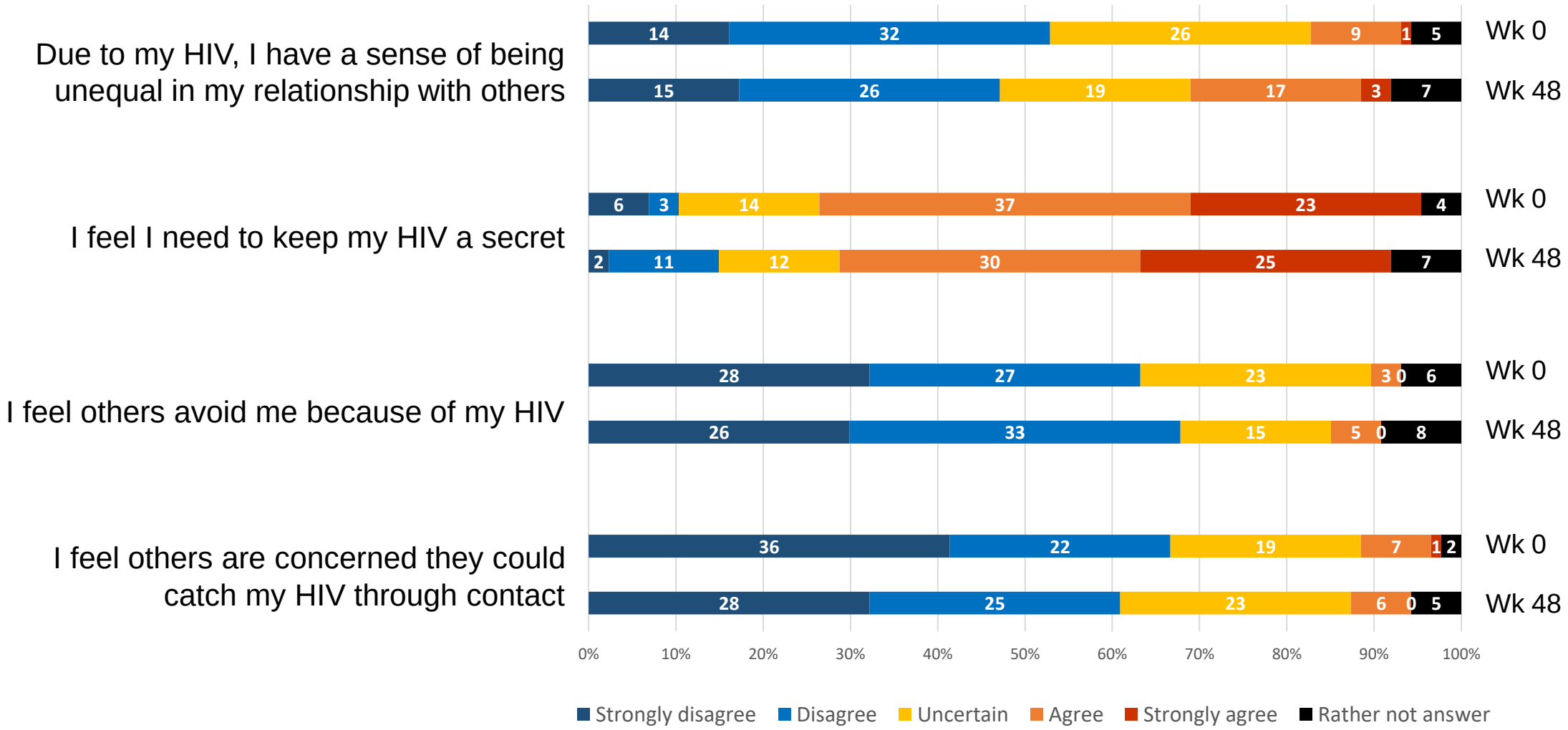
Applies to any contingency table

Observations must be mutually exclusive

carer_bl	group paHIV	HEU	HUU	Total
Parent	184	97	102	383
	200.2	85.6	97.2	383.0
	76.35	94.17	87.18	83.08
Grandparent	39	2	10	51
	26.7	11.4	12.9	51.0
	16.18	1.94	8.55	11.06
Aunt/Uncle	18	4	5	27
	14.1	6.0	6.9	27.0
	7.47	3.88	4.27	5.86
Total	241	103	117	461
	241.0	103.0	117.0	461.0
	100.00	100.00	100.00	100.00

Pearson chi2(4) = 19.4573 Pr = 0.001
Fisher's exact = 0.000

Presenting Likert scale data longitudinally



Compare 2 trials:

	VL ≤50c/mL	VL >50c/mL	Total
Arm A	11	25	36
Arm B	45	42	87
Total	56	67	123

Chi-square = 3.79
P = 0.052

	VL ≤50c/mL	VL >50c/mL	Total
Arm A	11	26	37
Arm B	45	41	86
Total	56	67	123

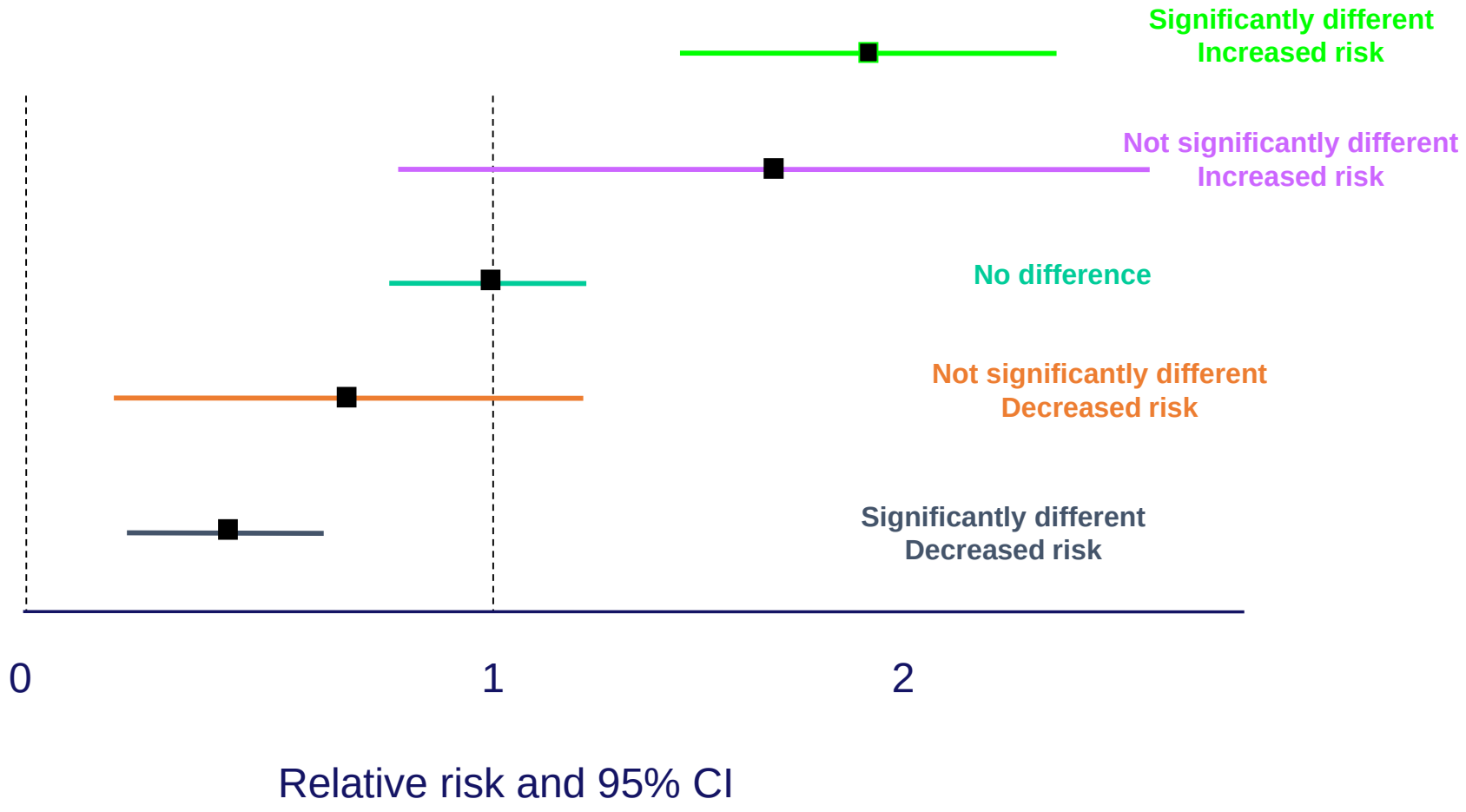
Chi-square = 4.45
P = 0.035

Small changes in the data can switch the significance of the results

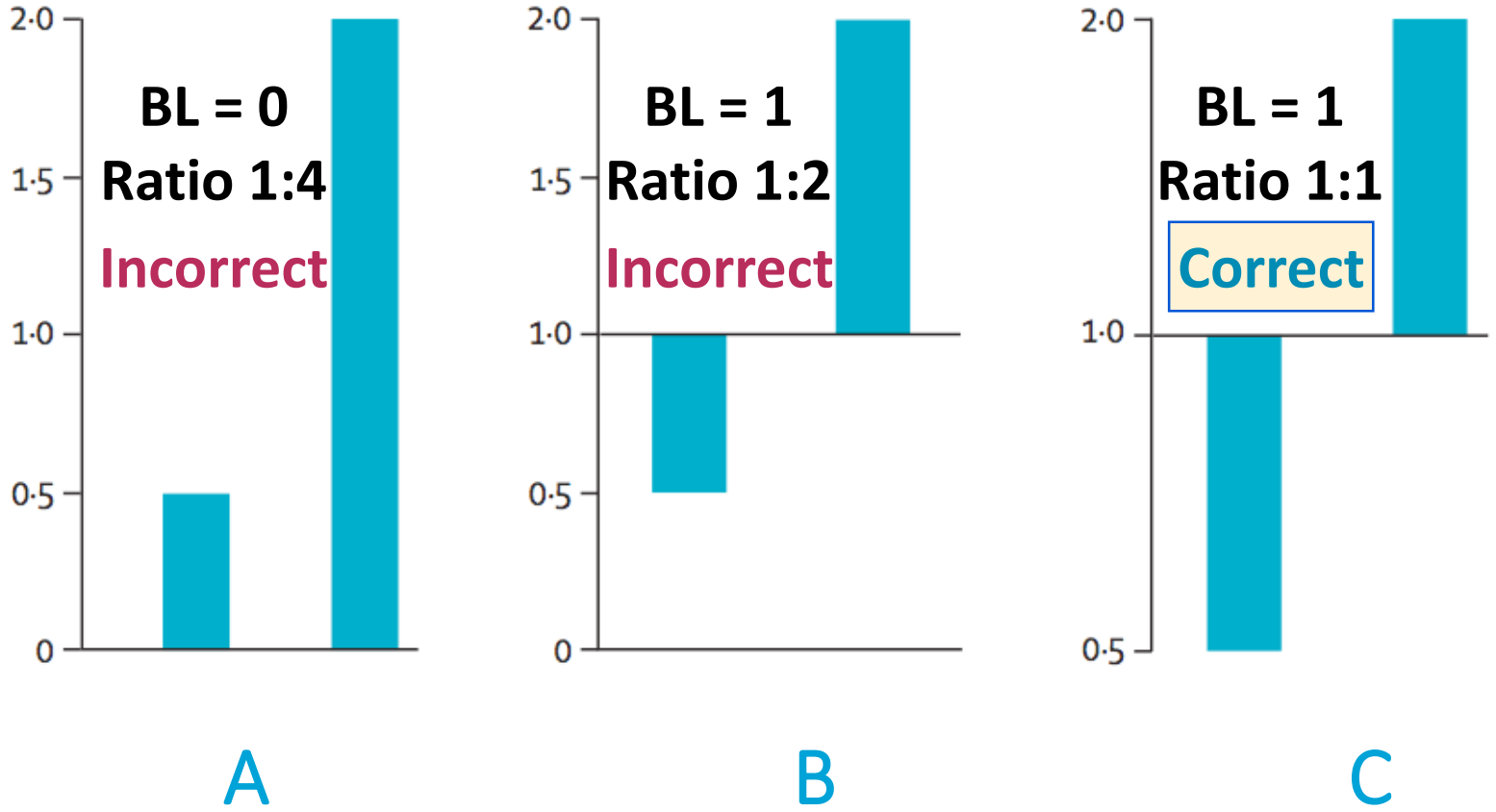
In large trials - results can be statistically significant, but not clinically important

Relative risk, or OR or HR

As a graphical representation



Symmetrical around 1 if you reverse the comparator



Graphical representation of risk ratios of 0.5 and 2, considering baseline and scale

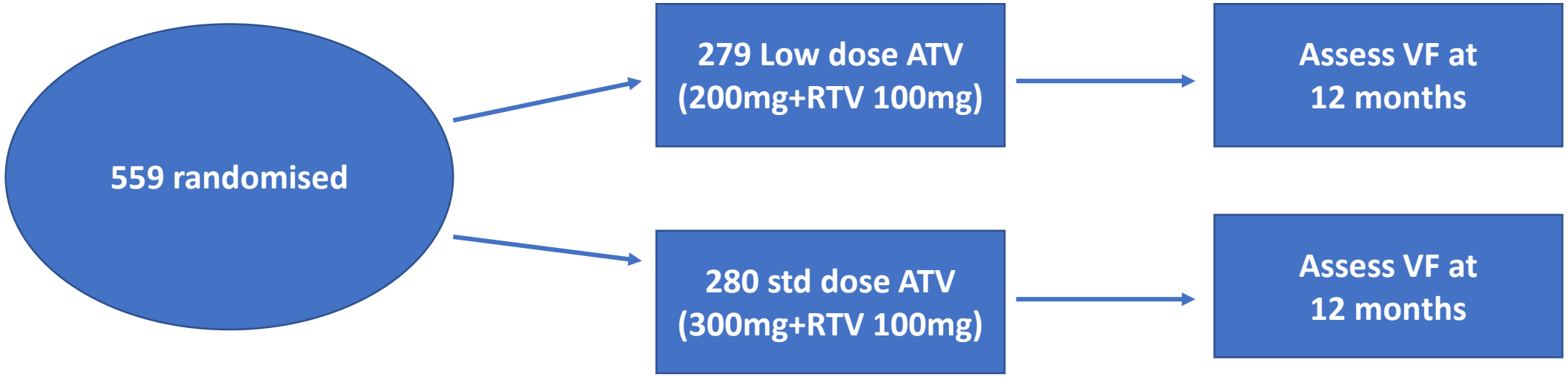
Proportions in RCT’s

- RCT analyses are usually unadjusted comparisons

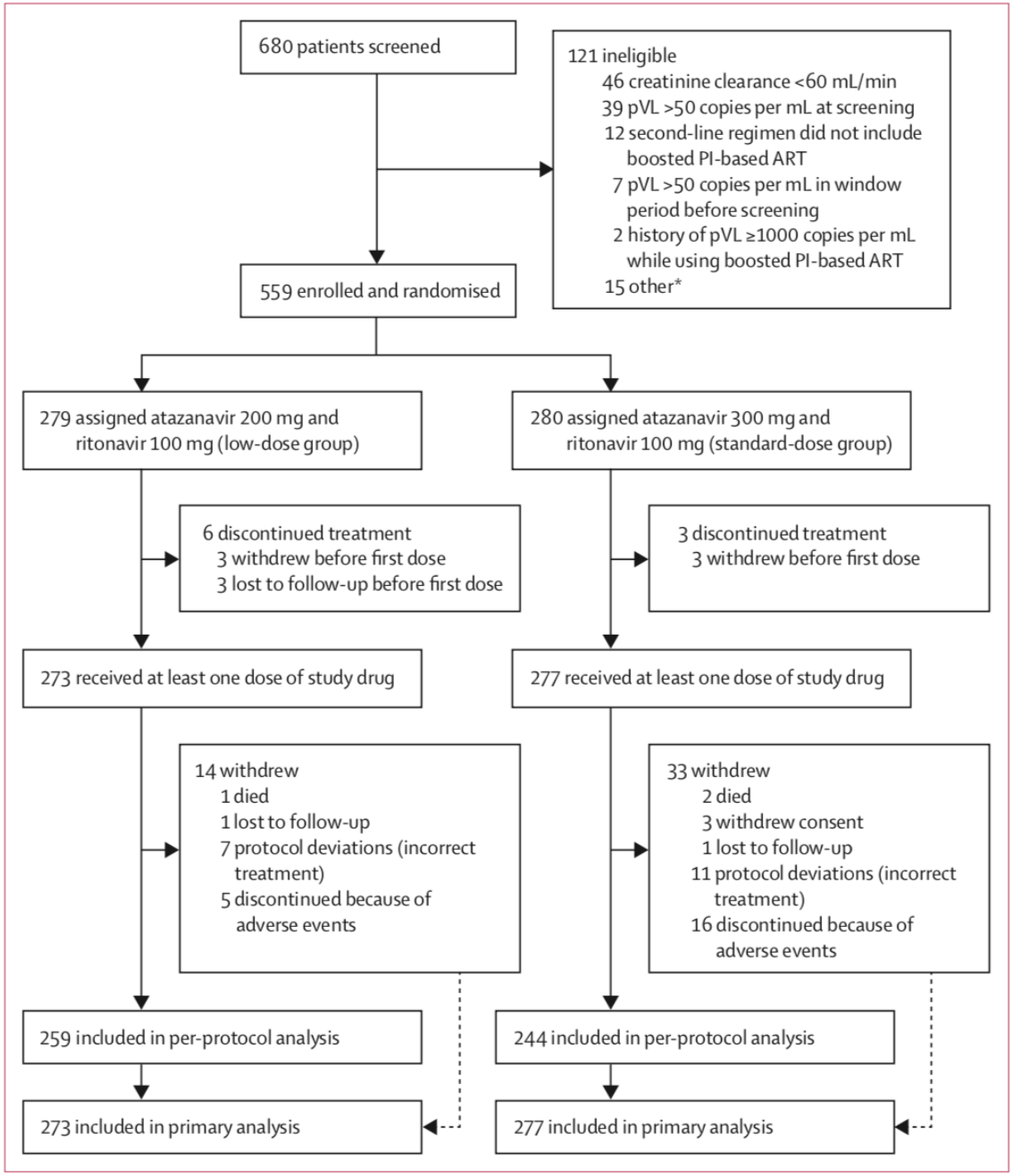
- Why??

Low-dose versus standard-dose ritonavir-boosted atazanavir in virologically suppressed Thai adults with HIV (LASA): a randomised, open-label, non-inferiority trial

Bunupuradah, T et al. Lancet HIV, 2016.



Non-inferiority – lower limit of the 95%CI for the absolute difference in virological response rates for the low dose vs standard dose arm would lie above -10%

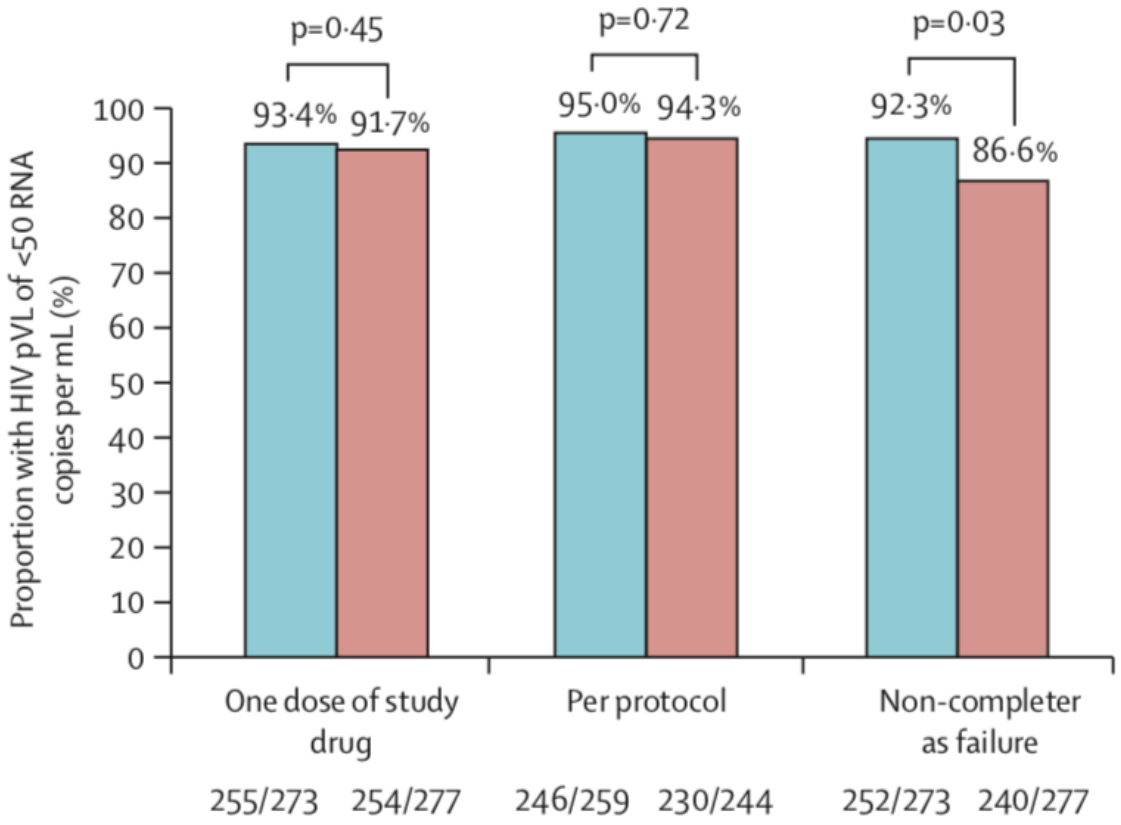

680 patients screened
121 ineligible
46 creatinine clearance <60 mL/min
39 pVL >50 copies per mL at screening
12 second-line regimen did not include boosted PI-based ART
7 pVL >50 copies per mL in window period before screening
2 history of pVL ≥1000 copies per mL while using boosted PI-based ART
15 other*
559 enrolled and randomised
279 assigned atazanavir 200 mg and ritonavir 100 mg (low-dose group)
280 assigned atazanavir 300 mg and ritonavir 100 mg (standard-dose group)
6 discontinued treatment
3 withdrew before first dose
3 lost to follow-up before first dose
3 discontinued treatment
3 withdrew before first dose
273 received at least one dose of study drug
277 received at least one dose of study drug
14 withdrew
1 died
1 lost to follow-up
7 protocol deviations (incorrect treatment)
5 discontinued because of adverse events
33 withdrew
2 died
3 withdrew consent
1 lost to follow-up
11 protocol deviations (incorrect treatment)
16 discontinued because of adverse events
259 included in per-protocol analysis
244 included in per-protocol analysis
273 included in primary analysis
277 included in primary analysis

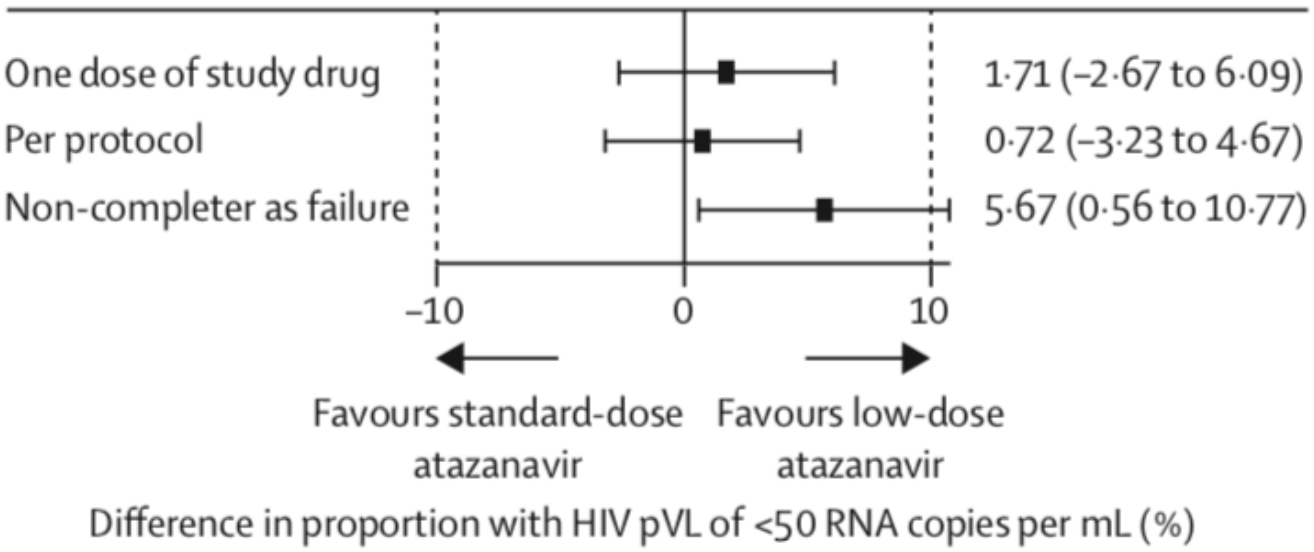
Demographic characteristics expressed as proportions

No p-values for comparison of demographics of randomised arms (Consort 2010 guidelines)

	Low-dose atazanavir (n=273)	Standard-dose atazanavir (n=277)	Total (n=550)
Age (years)	42 (7·6)	41 (7·4)	42 (7·5)
Male sex	129 (47%)	145 (52%)	274 (50%)
Bodyweight (kg)	59 (11)	60 (11)	59 (11)
Height (cm)	162 (8·4)	163 (8·8)	162 (8·6)
Body-mass index (kg/m²)	22·7 (3·5)	22·4 (3·4)	22·5 (3·5)
CD4 nadir (cells per µL)	120 (137)	126 (133)	123 (135)
History of ever used dual NRTIs	35 (13%)	39 (14%)	74 (13%)
Duration of treatment with boosted protease-inhibitor-containing regimen before screening (years)	5·2 (3·0)	5·1 (3·1)	5·2 (3·0)
Boosted protease inhibitor at screening (%)			
Lopinavir	233 (85%)	234 (84%)	467 (85%)
Indinavir	22 (8%)	24 (9%)	46 (8%)
Saquinavir	18 (7%)	19 (7%)	37 (7%)
NRTIs at screening (%)			
Lamivudine	223 (82%)	222 (80%)	445 (81%)
Tenofovir	203 (74%)	202 (73%)	405 (74%)
Zidovudine	108 (40%)	124 (45%)	232 (42%)
Didanosine	16 (6%)	14 (5%)	30 (5%)
Stavudine	10 (4%)	12 (4%)	22 (4%)
CD4 count (cells per µL)	549 (256)	528 (222)	539 (240)
HIV RNA concentration <50 copies per mL	263 (96%)	266 (96%)	529 (96%)

Results





Low-dose atazanavir Standard-dose atazanavir

Discontinuations – why is the NC=F analysis better?

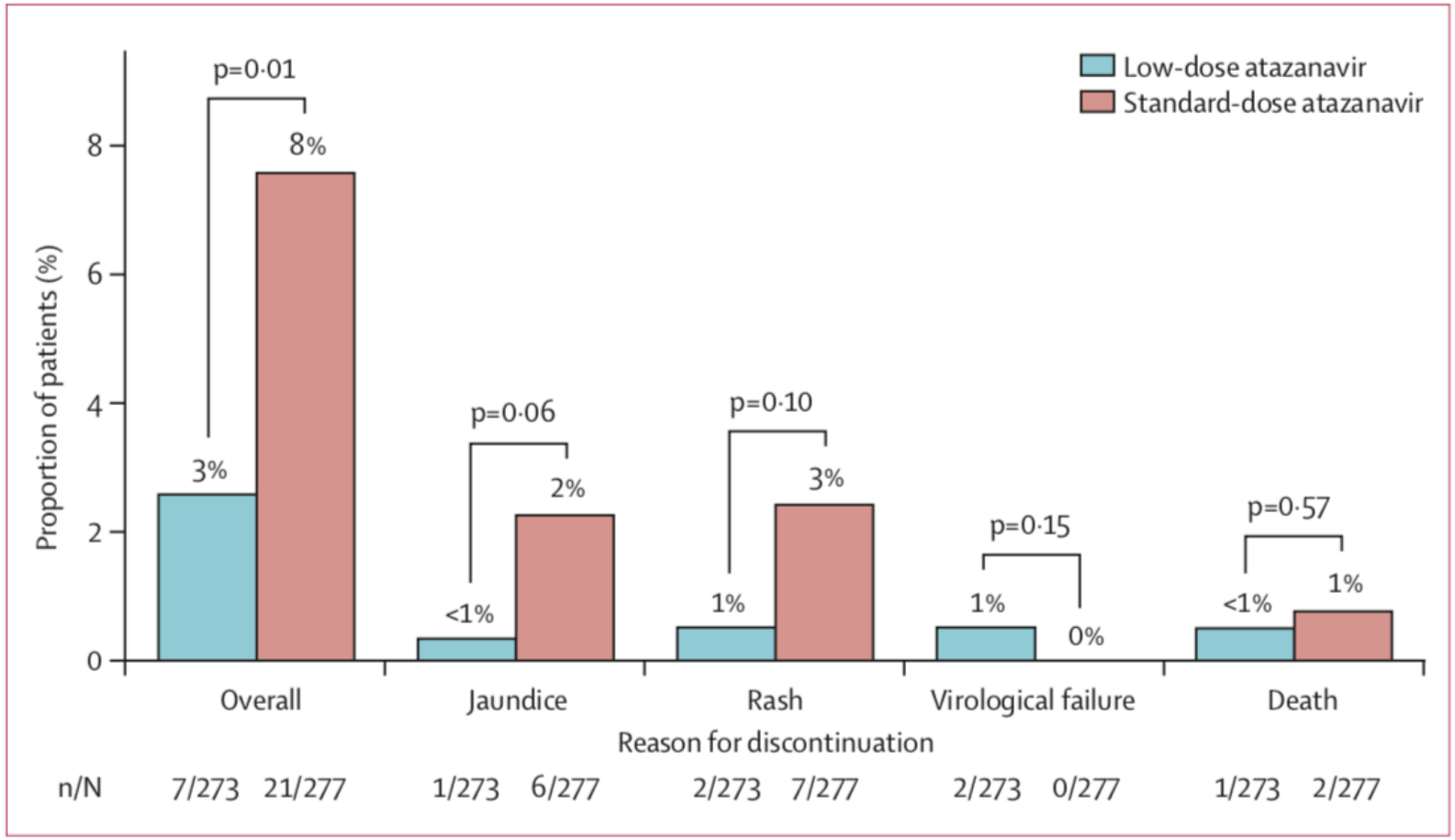


Figure 3: **Reasons for discontinuation**

Observational studies

- Cross-sectional
- Cohort
- Case control

Observational studies

- Cross-sectional (prevalence of a characteristics)
- Cohort (time to occurrence of a characteristic, repeated measurements of endpoints)
- Case control (investigation of rare diseases).

- Observational study analyses are usually adjusted analyses, because of the possibility of confounding

Case control

- To investigate a rare outcome
- The presence of the disease or characteristics is the outcome
- Factors associated with that outcome assessed using *conditional* logistic regression for matched studies (acknowledges the cases and controls are linked, and compares them in the same ‘set’).

- For a rare outcome (as in a case control study – the OR approximates the RR).

Odds vs. Risk=probability

If the risk is…	Then the odds are…
½ (0.5)	**1:1 (1)**
¾ (0.75)	**1:3 (0.33)**
1/10 (0.1)	**1:9 (0.111)**
1/20 (0.05)	**1:19 (0.053)**
1/100 (0.01)	**1:99 (0.0101)**

Note: An odds is always higher than its corresponding relative risk, unless the probability is 100%.

Cross-sectional studies

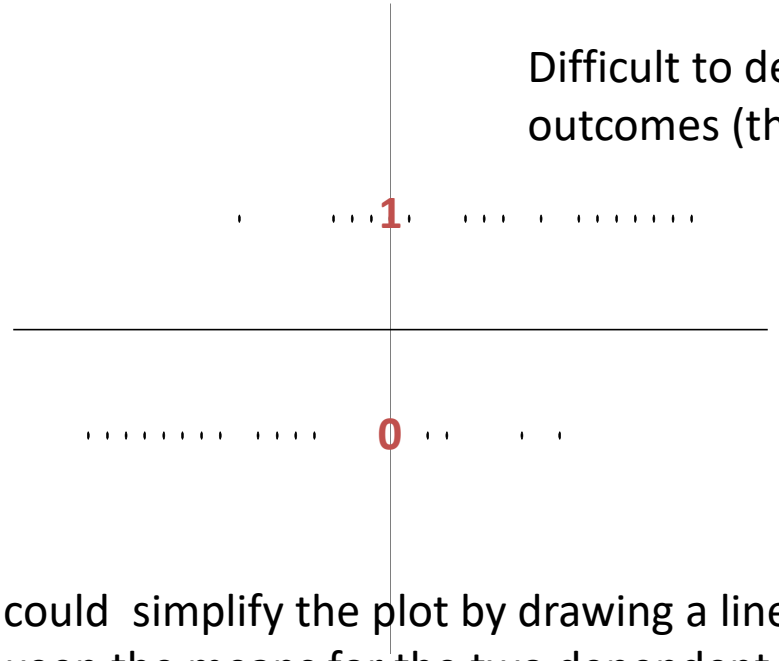
- **STROBE reporting guidelines for observational studies:**
  9. Describe any efforts to address potential sources of bias

- Because of the possibility of confounding and bias, we want to compare baseline characteristics by group, and adjust for these where there are differences, particularly in prognostic variables.

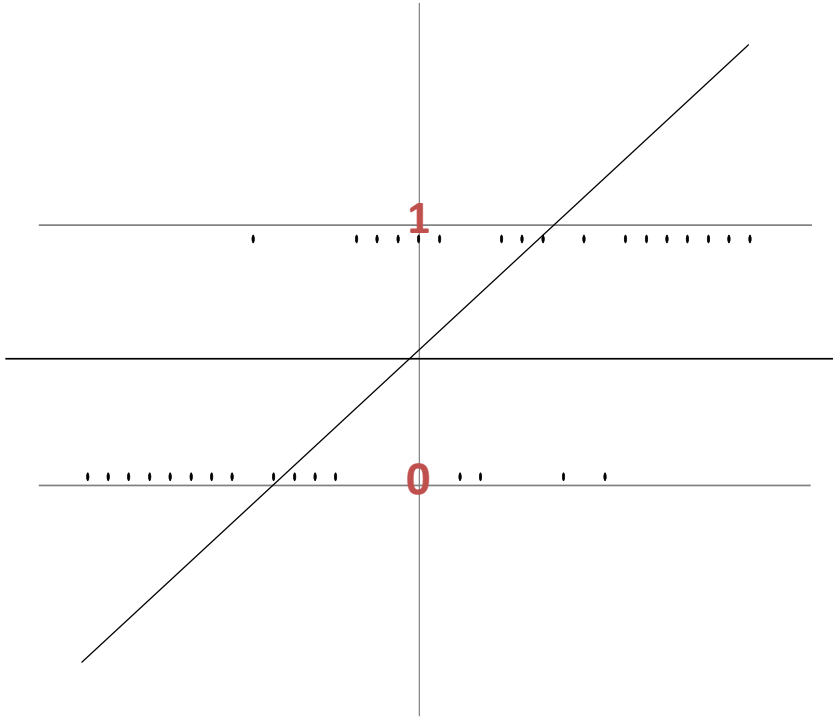
- When the outcome is prevalence (or a dichotomous variable), adjusted (multivariate) analyses are conducted using logistic regression

Outcome has 2 possible values (Yes = 1 or No = 0)

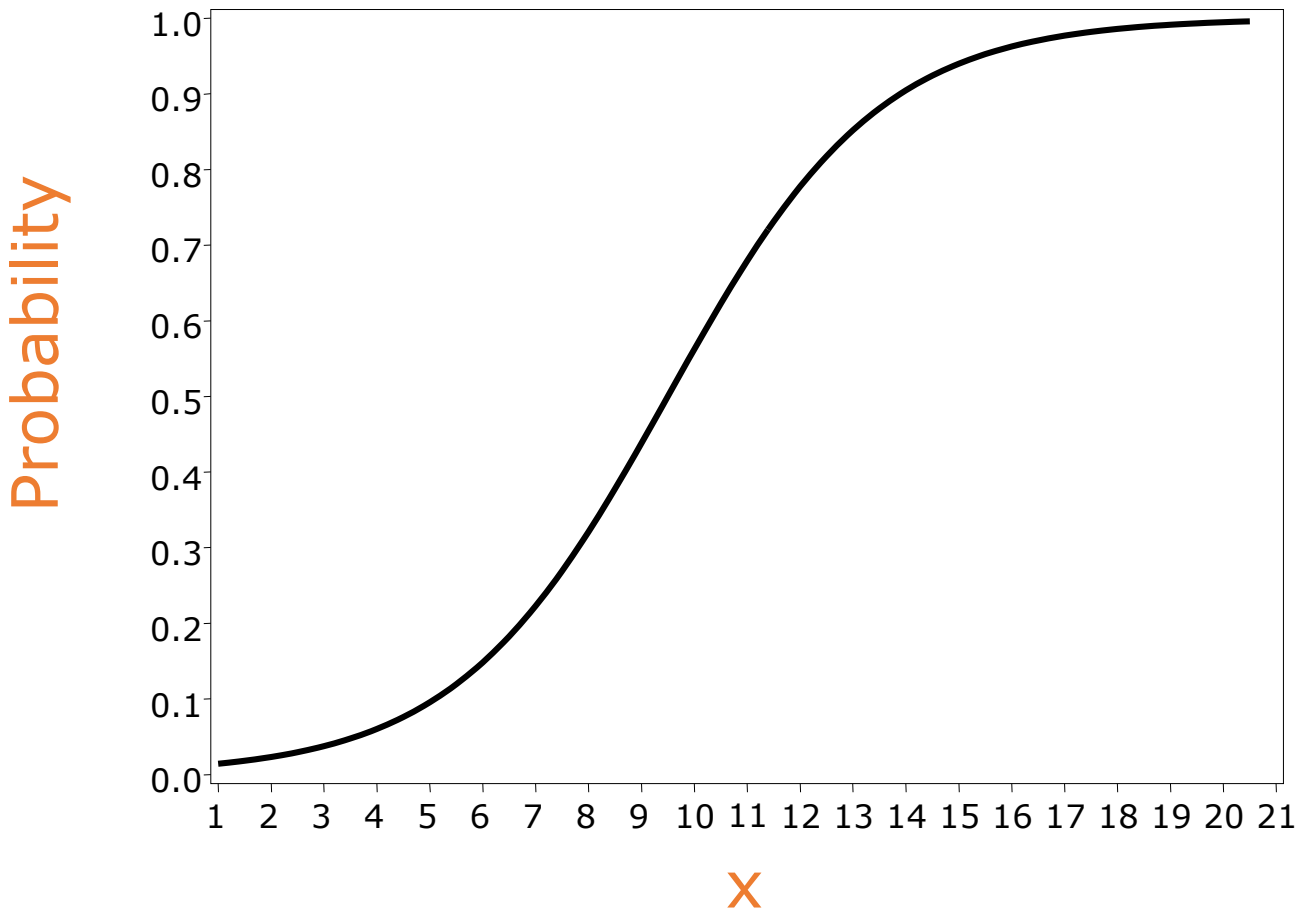
Difficult to decipher a relationship between the outcomes (the probability an event occurs, or not)



We could simplify the plot by drawing a line between the means for the two dependent variable levels, but (a) the line oversimplifies the relationship and (b) it gives predictions that cannot be observable values of Y for extreme values of X (ie > 1 or < 0).



Logistic Regression Curve

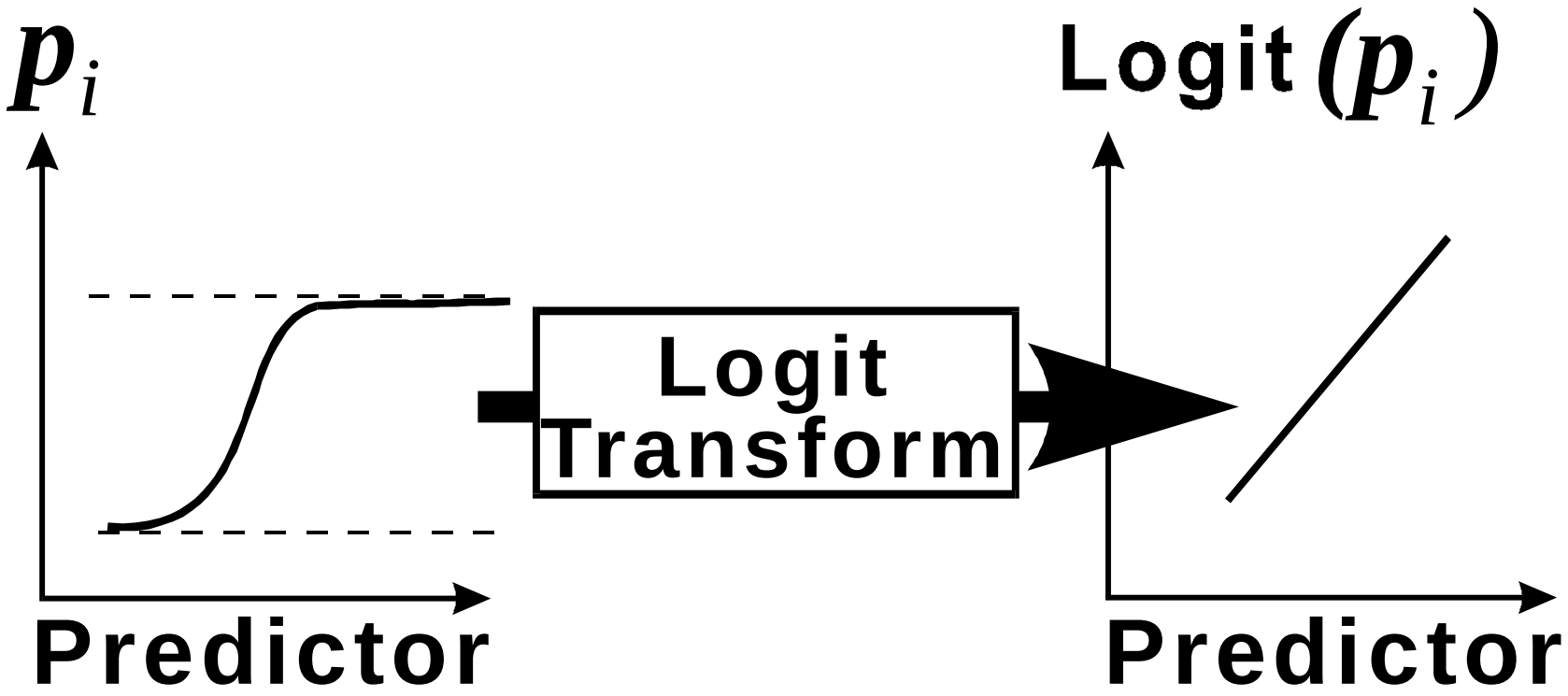


Logit Transformation

Logistic regression models transform probabilities called *logits*.

$$\text{logit}(\ p_i) \ = \ \log\left(\frac{p_i}{1-p_i}\right)$$

Assumption



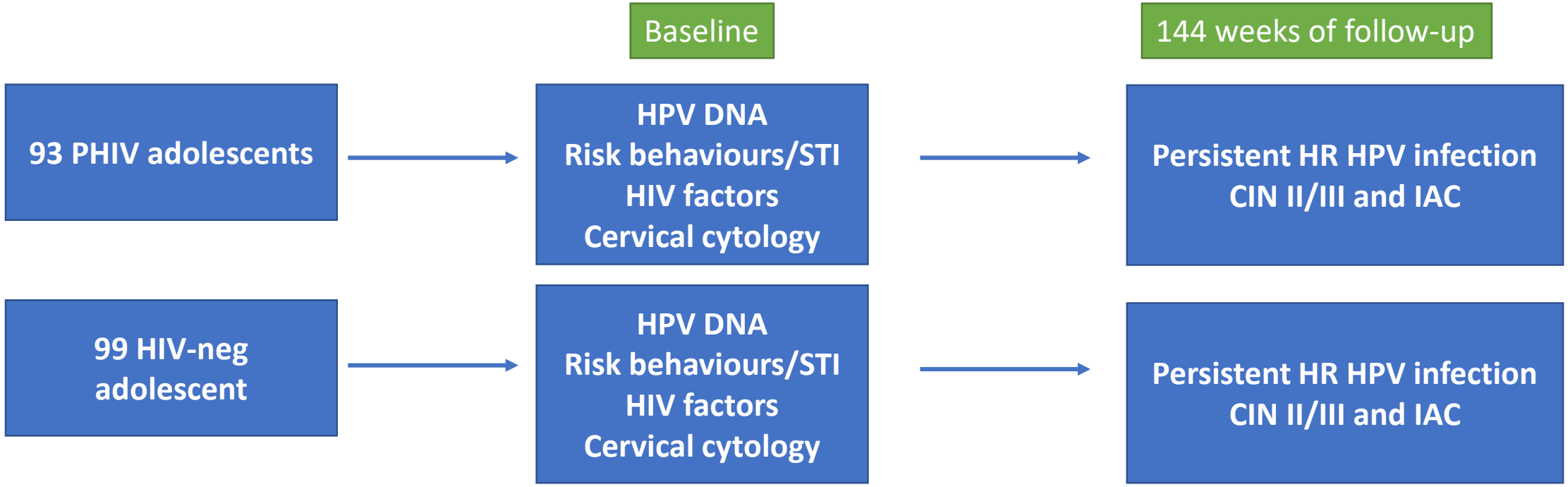
Parameter interpretation (ctd).

- Exponentiating both sides of the logit link function we get the following:

$$\left(\frac{p_i}{1-p_i}\right) = \text{odds} = e^{\beta 0} e^{\beta 1 X1}$$

Risk Factors for Human Papillomavirus Infection and Abnormal Cervical Cytology Among Perinatally Human Immunodeficiency Virus-Infected and Uninfected Asian Youth

Sohn, A et a. Clin Infect Dis. (2018). 67:606



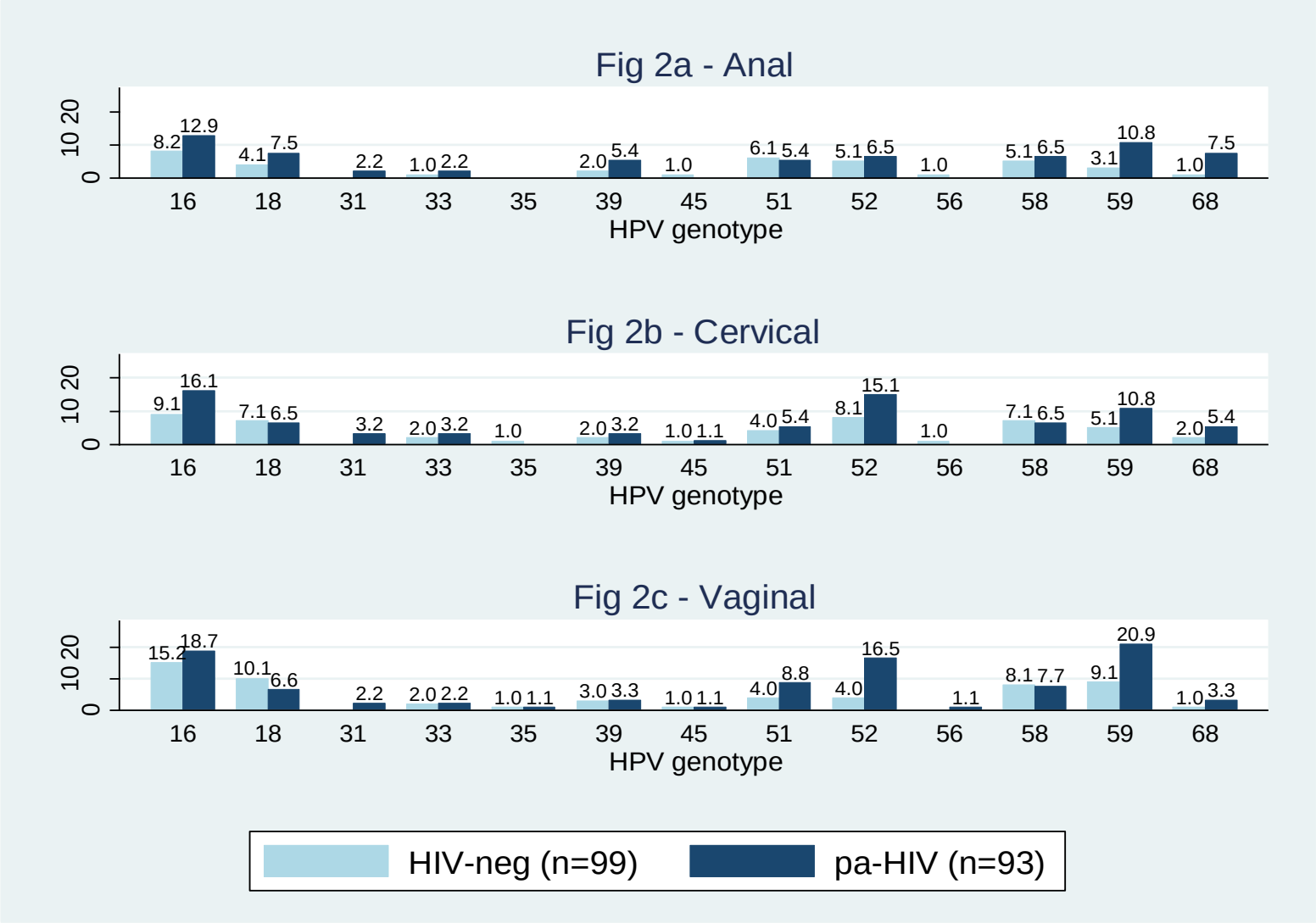
Characteristics at study entry

Characteristic	Total (n = 192)	Perinatally HIV Infected (n = 93)	HIV Uninfected (n = 99)	*P* Value
Age, years	19 (18–20)	19 (17–20)	19 (18–20)	.36
Living situation				<.001
One or both parents	84 (44%)	24 (26%)	60 (61%)	…
Relatives	44 (23%)	33 (35%)	11 (11%)	…
Partner	43 (22%)	24 (26%)	19 (19%)	…
Alone	15 (8%)	8 (9%)	7 (7%)	…
Other, did not answer	6 (3%)	4 (4%)	2 (2%)	…

Characteristics at baseline (cont'd)

Characteristic	Total (n = 192)	Perinatally HIV Infected (n = 93)	HIV Uninfected (n = 99)	*P* Value
Condom use with vaginal sex in past 6 months				<.001
Always	42 (22%)	31 (33%)	11 (11%)	…
Sometimes	93 (48%)	50 (54%)	43 (43%)	…
Never	52 (27%)	10 (11%)	42 (42%)	…
No recent sex	5 (3%)	2 (2%)	3 (3%)	…
Any anogenital HPV	133 (69%)	74 (80%)	59 (60%)	.003
Any anogenital HR HPV	99 (52%)	56 (60%)	43 (43%)	.02

Prevalence of HR HPV subtypes by site



Factors associated with any HR HPV

	UNIVARIATE		MULTIVARIATE	
Covariate	OR (95%CI)	P	OR (95%CI)	P
PHIV vs HIV-neg	1.97 (1.11 – 3.50)	0.02	2.02 (1.09 – 3.77)	0.03
Ever (vs never) been pregnant?	0.61 (0.34 – 1.09)	0.09	0.61 (0.32 – 1.15)	0.12
Asymptomatic STI (CT or NG)	2.72 (1.38 – 5.36)	0.003	1.99 (0.9 – 4.14)	0.06
Condom use in last 6 months		0.45		
Always or no sex	1 (ref)			
Sometimes/never	1.29 (0.67 – 2.49)			
Lifetime sexual partners		<0.001		<0.001
1	1 (ref)		1 (ref)	
≥2	3.90 (2.09 - 7.28)		3.46 (1.78 – 6.71)	
Drug use in past 3 months	1.91 (0.46 – 7.89)	0.36		

Describing the results in text

- **Association Between Participant Demographic and Behavioral Characteristics and HR-HPV in Anogenital Compartments**
- We used logistic regression to assess associations between participant characteristics and behaviors and any detected HR-HPV infection in the cervix, vagina, or anus (Table 2). In a univariate analysis, HIV status, increasing number of life- time partners, and having an asymptomatic STI diagnosed at enrollment were associated with a significantly higher odds of having HR-HPV, while ever being pregnant was associated with a reduced odds. After adjusting for ever having been pregnant and having an asymptomatic STI at enrollment in a multivariate model, PHIV participants were more likely to have HR-HPV than HU adolescents (odds ratio [OR], 2.02; 95% confidence interval [CI], 1.09–3.77; $P = .03$). Overall, compared to those with 1 lifetime partner, those with ≥2 partners had an increased odds of HR-HPV infection (OR, 3.46; 95% CI, 1.78–6.71; $P < .001$).

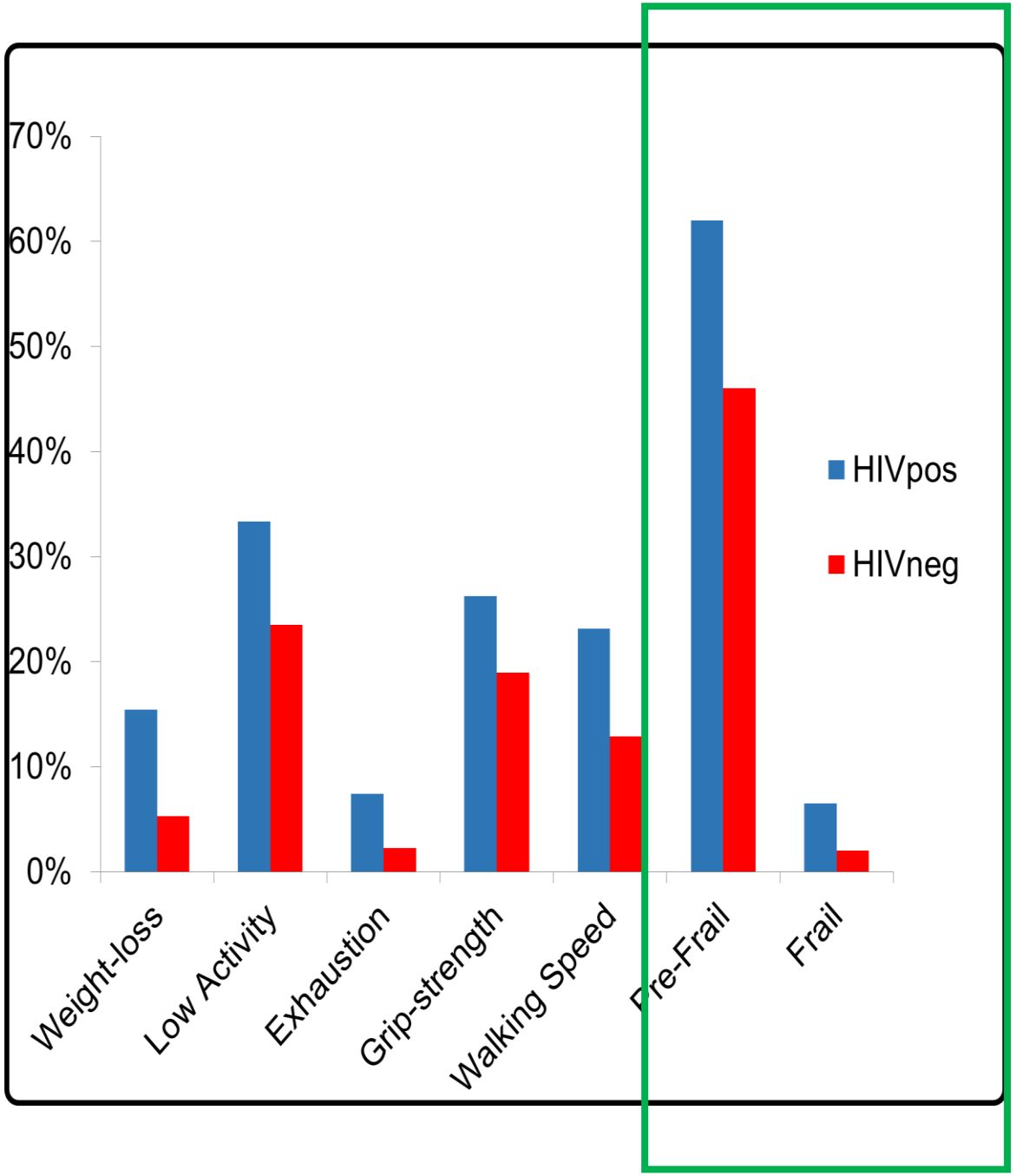
How do I decide on terms in a multivariate model?

- What sort of model do I want?
  - Predictive or explanatory?

- Univariate models:
  - consider biological plausibility
  - consider the role of known confounders

Conservative rule: adjust for co-variates with a P in univariate models of $P<0.2$

More realistic rule: ajdust for for co-variates with a univariate P of $P<0.1$



Ordinal or multinomial logistic?

3 outcomes:

1) Not frail
2) Pre-frail
3) Frail