Introduction to Medical Statistics 2026

Oxford University Clinical Research Unit
March 23-27, 2026

Ronald Geskus and the biostatistics crew
Oxford University Clinical Research Unit
Hospital for Tropical Diseases,
Ho Chi Minh City, Viet Nam

Part X

Analysis and Reporting

Contents:

1. Which variables to include
2. How to include variables
3. Hypothesis testing and statistical significance
4. Reporting
5. Help!

Criteria for reviewers (PLOS neglected tropical diseases)

Methods:

- Are the objectives of the study clearly articulated with a clear testable hypothesis stated?
- Is the study design appropriate to address the stated objectives?
- Is the population clearly described and appropriate for the hypothesis being tested?
- Is the sample size sufficient to ensure adequate power to address the hypothesis being tested?
- Were correct statistical analysis used to support conclusions?
- Are there concerns about ethical or regulatory requirements being met?

Criteria for reviewers (PLOS neglected tropical diseases)

Results:

- Does the analysis presented match the analysis plan?
- Are the results clearly and completely presented?
- Are the figures (Tables, Images) of sufficient quality for clarity?

Conclusions:

- Are the conclusions supported by the data presented?
- Are the limitations of analysis clearly described?
- Do the authors discuss how these data can be helpful to advance our understanding of the topic under study?
- Is public health relevance addressed?

Outline

Which variables to include in our regression model
- Variable reduction
- Multiple testing and fishing expeditions

How to include variables
- Stratified analysis
- Dichotomania

The role of p-values
- Examples

Reporting

Help!

oucru

UNIVERSITY OF OXFORD

wellcome

Which variables to include?

- Use expert knowledge

Which variables to include?

- Use expert knowledge
- Inferential: any variable that is possibly related to the outcome

Which variables to include?

- Use expert knowledge
- Inferential: any variable that is possibly related to the outcome
- Predictive: any variable that helps better predicting/diagnosing the outcome
  - does not need to have a direct causal relation with outcome
  - proxy variable that is easy and cheap to measure may be preferred

Which variables to include?

- Use expert knowledge
- Inferential: any variable that is possibly related to the outcome
- Predictive: any variable that helps better predicting/diagnosing the outcome
  - does not need to have a direct causal relation with outcome
  - proxy variable that is easy and cheap to measure may be preferred
- Causal
  - observational design: correct for confounders to reduce, or ideally eliminate, bias
  - randomized design: no additional variables need to be included to reduce bias.
    However, power can be increased by including independent baseline variables that are strongly related to the outcome
    - may explain part of the variation in outcome
    - there may be random imbalance (especially in small trials)

Dexamethasone in Hospitalized Patients with Covid-19

Recovery Trial; imbalance in important predictor of outcome.

Through the play of chance in the unstratified randomization, the mean age was 1.1 years older among patients in the dexamethasone group than among those in the usual care group. To account for this imbalance in an important prognostic factor, estimates of rate ratios were adjusted for the baseline age in three categories ($<$70 years, 70 to 79 years, and $\geq$ 80 years).

Outline

Which variables to include in our regression model
- Variable reduction
- Multiple testing and fishing expeditions

How to include variables
- Stratified analysis
- Dichotomania

The role of p-values
- Examples

Reporting

Help!

oucru

UNIVERSITY OF OXFORD

wellcome

Stepwise regression

- Forward
  1. Fit regression models for each variable separately. Select the one with smallest p-value
  2. Fit regression models with two variables: the one selected in 1. and one-by-one each of the other ones. Select the one with smallest p-value amongst the other ones
  3. Repeat this until all variables not in the model are not significant when added
- Backward
  - Start with model including all variables
  - Eliminate one-by-one based on largest p-value until all p-values smaller than some value
- Combination: include all variables from univariate analyses with p-value < 0.1 (or some other number) in multivariable model and perform backward regression until all $p < 0.05$

Stepwise regression and study question

- Inferential
  - often parsimonious model chosen that only includes statistically significant variables
  - ok to obtain first idea of relationships, but results may be hard to interpret
  - example: risk factors for STI's in sex workers in China

Table 2. Multivariate analyses of the relationship between general and sexual characteristics, HIV/sexually transmitted disease knowledge and self-efficacy and sexually transmitted diseases among 966 sex workers at Guangzhou, China, March 1998–October 1999.

	HIV OR (95% CI)	Gonorrhoea OR (95% CI)	Trichomoniasis OR (95% CI)	Chlamydia OR (95% CI)	Syphilis[a] OR (95% CI)
General					
Age (years)					
< 21				2.0 (1.2–3.4)	
21–22				1.2 (0.7–2.2)	
23–25				1.4 (0.8–2.5)	NS (–)
26–30				1.0 (0.6–1.7)	
> 30				1	
Not always lived in Guangzhou	0.2 (0.0–0.8)			NS (+)	NS (–)
Injected drugs (since 1990)	8.0 (2.1–30.3)		2.5 (1.3–5.0)	0.2 (0.1–0.4)	NS (+)
Recruitment clients on street/via pimps	NS (+)	NS (+)	NS (+)		2.6 (1.7–4.0)
No regular salary	NS (+)	2.3 (1.3–4.1)	6.1 (3.5–10.4)		
STD check-up (past 12 months)		0.4 (0.2–0.9)		NS (–)	NS (–)
Sexual activity					
Number of clients (per week)					
< 6	1				1
6–7	0.1 (0.0–0.7)	NS (+)	NS (+)		1.7 (1.0–3.0)
> 7	0.2 (0.0–1.2)				2.2 (1.3–3.7)
Duration of prostitution (years)					
< 1					1
1				NS (–)	1.5 (0.9–2.7)
3					2.5 (1.4–4.6)
> 3					1.8 (0.9–3.4)
Steady partner (past 12 months)					
No steady partner					1
1 steady partner			NS (–)		2.1 (1.4–3.2)
> 1 steady partner					1.1 (0.6–2.3)
Knowledge					
Knowledge about AIDS		NS (–)	NS (–)	NS (–)	0.9 (0.8–1.0)
Knowledge about condom use		NS (–)	0.9 (0.7–1.0)	0.8 (0.8–0.9)	NS (–)
Condom use during vaginal sex (past 2 months)					
Always		1		1	1
Frequently		3.1 (3.2–8.3)		2.1 (1.0–4.9)	1.5 (0.9–2.6)
Rarely		8.6 (3.2–23.3)	NS (+)	2.2 (0.7–2.2)	2.5 (1.3–4.6)
Never		9.6 (3.0–30.4)		2.7 (1.5–7.3)	3.8 (1.7–8.7)
Diagnostic evidence of:					
HIV			5.0 (1.4–17.0)		5.7 (1.6–20.7)
Gonorhoea			2.3 (1.3–4.1)	2.9 (1.7–4.7)	
Trichomoniasis	11.2 (2.9–42.7)	1.8 (1.0–3.3)		2.6 (1.6–4.1)	NS (+)
Chlamydia		3.0 (1.8–5.0)	2.8 (1.7–4.4)		

CI, Confidence interval; OR, odds ratio; STD, sexually transmitted diseases.
In building multivariate models, all univariate predictors (with P value < 0.10) were included in a stepwise backward procedure. All overall P values included in the multivariate models and presented in the table are ≤ 0.05.
A blank cell indicates that the variable was not included in the multivariate model ($P < 0.10$ in univariate analyses).
NS indicates that the variable after inclusion in the multivariate model was no longer statistically significant, although in univariate analyses the model P value was < 0.10. The (+) or (–) behind the 'NS' indicates the direction of the risk estimate in univariate analyses.
[a] *Treponema pallidum* haemagglutination assay positive.

Stepwise regression and study question

- Inferential
  - often parsimonious model chosen that only includes statistically significant variables
  - ok to obtain first idea of relationships, but results may be hard to interpret
  - example: risk factors for STI's in sex workers in China
- Predictive: ok, but validation very important

Stepwise regression and study question

- Inferential
  - often parsimonious model chosen that only includes statistically significant variables
  - ok to obtain first idea of relationships, but results may be hard to interpret
  - example: risk factors for STI's in sex workers in China
- Predictive: ok, but validation very important
- Causal: not recommended
  - Choice should not be determined by p-value, but by veriable being confounder or having strong relationship with outcome

Drawbacks stepwise regression in inferential research

- Excluded covariables may in reality have effect (power):
  “absence of proof” is not “proof of absence”
  Small effects would be included if sample were much larger
- Repeated testing, such that risk of spurious significant results increases
- Based on methods that were intended to be used to test prespecified hypotheses

Drawbacks stepwise regression in inferential research

- Excluded covariables may in reality have effect (power):
  "absence of proof" is not "proof of absence"
  Small effects would be included if sample were much larger
- Repeated testing, such that risk of spurious significant results increases
- Based on methods that were intended to be used to test prespecified hypotheses
- Stepwise methods are often a complicated equivalent to throwing darts blindfolded (the final model is more due to random chance than anything else)
- See also What are some of the problems with stepwise regression?

Outline

Which variables to include in our regression model

- Variable reduction
- Multiple testing and fishing expeditions

How to include variables

- Stratified analysis
- Dichotomania

The role of p-values

- Examples

Reporting

Help!

oucru

UNIVERSITY OF OXFORD

wellcome

Example: Aspirin trial (Lancet 1988; 2: 349–60)

- A RCT in patients with a cardiac infarction showed a highly significant survival benefit of aspirin therapy vs. placebo
- Prior to publication, the editors of the Lancet asked for 40 additional subgroup analyses
- The authors agreed under the condition that they can chose any additional characteristic themselves

Example: Aspirin trial (II)

- The authors chose the zodiac sign
- Gemini and libra showed a slightly higher mortality in the aspirin arm compared to placebo
- For all other zodiac signs, aspirin was highly significantly superior
- "All these subgroup analyses should, perhaps, be taken less as evidence about who benefits than as evidence that such analyses are potentially misleading"

What's the problem?

- Assume we perform significance tests of N independent null hypotheses which are all true
- Probability that a specific null hypothesis is falsely rejected: 0.05
- Probability that at least one of the null hypotheses is falsely rejected:
  $1 - P(\text{none rejected}) = 1 - 0.95^N$

N	5	10	20	50
$1 - 0.95^N$	23%	40%	64%	92%

- Similar problems occur for general (possibly dependent) multiple significance tests

Instances of multiple testing

- Multiple testing often occurs
  - Subgroup analyses
  - Several endpoints
  - Pairwise comparisons of $>$ 2 subgroups
  - Interim analyses
  - Data dredging

Instances of multiple testing

- Multiple testing often occurs
  - Subgroup analyses
  - Several endpoints
  - Pairwise comparisons of > 2 subgroups
  - Interim analyses
  - Data dredging
- What to do to cope with multiple testing
  - Pre-define primary study hypothesis and all important secondary hypotheses
  - Avoid unplanned fishing expeditions for significant results
  - Separate confirmatory and exploratory/inferential analyses
    - You cannot generate a hypothesis and statistically "prove" it based on the same data
  - May use statistical adjustments for multiple testing
    - Most simple (but very conservative): Bonferroni adjustment
    - Slightly better: Holm correction

Outline

Which variables to include in our regression model
- Variable reduction
- Multiple testing and fishing expeditions

How to include variables
- Stratified analysis
- Dichotomania

The role of p-values
- Examples

Reporting

Help!

Outline

Which variables to include in our regression model
- Variable reduction
- Multiple testing and fishing expeditions

How to include variables
- Stratified analysis
- Dichotomania

The role of p-values
- Examples

Reporting

Help!

oucru

UNIVERSITY OF OXFORD

wellcome

Example stratified analysis

HIV Prevalence and Associated Risk Factors among Individuals Aged 13-34 Years in Rural Western Kenya

Table 2. Age-group adjusted risk factors associated with HIV infection by gender among sexually active participants, N = 1202.

	Females					Males				
	N	Weighted HIV prevalence (%)	Age group aOR[a]	95% CI[b]	*P-value*	N	Weighted HIV prevalence (%)	Age group aOR	95% CI	*P-value*
Total	619	27.7	4.2	[2.7, 6.6]	<0.001	583	14.1	11.8	[6.4, 21.6]	<0.001
Age (years), overall median = 21	median = 22					median = 20				
IQR[c] = [18,28]	IQR = [18,28]					IQR = [17,26]				
Demographic characteristics										
Education										
Some primary school	316	30.1	*ref*[d]	–	NS[e]	230	9.5	*ref*	–	NS
Completed primary school	186	27.6	0.8	[0.5, 1.0]		190	17.6	1.1	[0.6, 2.0]	
Beyond primary school	117	22.2	0.6	[0.5, 1.2]		163	16.4	1.1	[0.7, 2.4]	
Occupation										
Has cash income[f]	213	40.7	2.1	[1.4, 3.1]	<0.001	215	22.3	1.2	[0.7, 2.0]	NS
Religion										
Muslim groups/other	117	36.2	*ref*	–	NS	109	11.2	*ref*	–	NS
Protestant groups	374	25.9	0.8	[0.4, 1.4]		321	16.4	1.6	[0.9, 3.0]	
Catholic groups	128	25.2	0.8	[0.5, 1.2]		153	11.1	1.0	[0.5, 2.1]	
Marital status										
Never married	224	8.6	*ref*		<0.001	369	5.6	*ref*	–	<0.05
Currently married	328	29.4	3.4	[1.5, 7.7]		189	32.4	2.4	[1.1,5.3]	
Divorced/Separated	6	36.2	5.0	[0.8, 21.0]		16	31.3	1.7	[0.5, 5.9]	
Widow/Widower	61	77.8	28.5	[10.6, 76.5]		9	49.7	5.2	[1.4, 19.7]	

oucru

UNIVERSITY OF OXFORD

wellcome

Why do a stratified analysis?

Why do a stratified analysis?

Because it is easier.

Why do a stratified analysis?

Because it is easier.

But,

same and more can be obtained by including stratum variable (e.g. gender) in one overall model:

- If relationship of variable with outome differs by gender:
  - Include interaction between gender and variable
  - Advantage: we can test and obtain p-value for interaction term
- If relationship of variable with outome does not differ by gender:
  - we can restrict parameter to be equal for both genders
  - advantage: more power

Outline

oucru

UNIVERSITY OF OXFORD

wellcome

Dichotomania

http://www.senns.uk/Geep.htm

Dichotomania is an obsessive compulsive disorder to which medical advisors in particular are prone. . . Show a medical advisor some continuous measurements and he or she immediately wonders. "Hmm, how can I make these clinically meaningful? Where can I cut them in two? What ludicrous side conditions can I impose on this?"

Dichotomization of continuous information tends to encourage dichotomous thinking, this limits the research questions we can ask and the conclusions we're able to draw. There is a literary canon decrying this practice written by statisticians (unfortunately it seems it is only read by statisticians).

Problems Caused by Categorizing Continuous Variables

- It assumes the relationship between the predictor and the response
  - is flat within intervals
  - abruptly changes as interval boundaries are crossed
- Researchers seldom agree on the choice of cutpoint. Some may compare blood pressure > 140 with ≤ 140, while others compare > 120 with ≤ 120.
- A patient does not report to her physician "my blood pressure exceeds 140" but rather reports 142 mmHg. The risk of stroke for this subject will be much lower than that of a subject with a blood pressure of 200 mmHg.
- Loss of power and precision

Outline

oucru

UNIVERSITY OF OXFORD

wellcome

When and how to use p-values

- P-value linked to hypothesis about population characteristic
  - Do not use p-value in baseline table

Do not use p-values in baseline table

Table 1. Demographic Data and Postoperative Rescue Medications

	Group D (n = 30)	Group DP (n = 30)	P value
Sex (M/F)*	16/14	13/17	0.438
Age (yr)†	8.1 ± 3.4	10.0 ± 3.9	0.052
Wt (kg)†	27.3 ± 11.5	34.7 ± 15.9	0.042
Ht (cm)†	127.5 ± 21.4	133.7 ± 20.5	0.257
Op. duration (min)†	119.3 ± 19.3	113.0 ± 26.9	0.299
Anes. duration (min)†	161.3 ± 23.3	165.1 ± 26.5	0.554
Recovery time (min)†	14.3 ± 8.4	12.1 ± 10.5	0.382
Postoperative analgesics/antiemetic			
Fentanyl consumption (μg/kg)†	10.7 ± 2.6	11.1 ± 2.0	0.534
Rescue analgesic needed (no. of patients)‡	14 (46.7%)	8 (26.7%)	0.180
Rescue antiemetic needed (no. of patients)‡	14 (46.7%)	13 (43.3%)	1.0

Data are mean ± SDM unless addressed. Statistical analyses were performed using *Chi-square test, †Student t-test, or ‡Fisher's exact test. Group D and DP represent dexamethasone only treated- and dexamethasone and propofol treated- patients, respectively.

- http://dx.doi.org/10.4097/kjae.2013.64.2.127

Do not use p-values in baseline table

Table 1. Demographic Data and Postoperative Rescue Medications

	Group D (n = 30)	Group DP (n = 30)	P value
Sex (M/F)*	16/14	13/17	0.438
Age (yr)†	8.1 ± 3.4	10.0 ± 3.9	0.052
Wt (kg)†	27.3 ± 11.5	34.7 ± 15.9	0.042
Ht (cm)†	127.5 ± 21.4	133.7 ± 20.5	0.257
Op. duration (min)†	119.3 ± 19.3	113.0 ± 26.9	0.299
Anes. duration (min)†	161.3 ± 23.3	165.1 ± 26.5	0.554
Recovery time (min)†	14.3 ± 8.4	12.1 ± 10.5	0.382
Postoperative analgesics/antiemetic			
Fentanyl consumption (μg/kg)†	10.7 ± 2.6	11.1 ± 2.0	0.534
Rescue analgesic needed (no. of patients)‡	14 (46.7%)	8 (26.7%)	0.180
Rescue antiemetic needed (no. of patients)‡	14 (46.7%)	13 (43.3%)	1.0

Data are mean ± SDM unless addressed. Statistical analyses were performed using *Chi-square test, †Student t-test, or ‡Fisher's exact test. Group D and DP represent dexamethasone only treated- and dexamethasone and propofol treated- patients, respectively.

- http://dx.doi.org/10.4097/kjae.2013.64.2.127
- What is the null hypothesis?
  There is no larger population that this table refers to

oucru

UNIVERSITY OF OXFORD

wellcome

Do not use p-values in baseline table

Table 1. Demographic Data and Postoperative Rescue Medications

	Group D (n = 30)	Group DP (n = 30)	P value
Sex (M/F)*	16/14	13/17	0.438
Age (yr)†	8.1 ± 3.4	10.0 ± 3.9	0.052
Wt (kg)†	27.3 ± 11.5	34.7 ± 15.9	0.042
Ht (cm)†	127.5 ± 21.4	133.7 ± 20.5	0.257
Op. duration (min)†	119.3 ± 19.3	113.0 ± 26.9	0.299
Anes. duration (min)†	161.3 ± 23.3	165.1 ± 26.5	0.554
Recovery time (min)†	14.3 ± 8.4	12.1 ± 10.5	0.382
Postoperative analgesics/antiemetic			
Fentanyl consumption (μg/kg)†	10.7 ± 2.6	11.1 ± 2.0	0.534
Rescue analgesic needed (no. of patients)‡	14 (46.7%)	8 (26.7%)	0.180
Rescue antiemetic needed (no. of patients)‡	14 (46.7%)	13 (43.3%)	1.0

Data are mean ± SDM unless addressed. Statistical analyses were performed using *Chi-square test, †Student t-test, or ‡Fisher's exact test. Group D and DP represent dexamethasone only treated- and dexamethasone and propofol treated- patients, respectively.

- http://dx.doi.org/10.4097/kjae.2013.64.2.127
- What is the null hypothesis?
  There is no larger population that this table refers to
- P-values make no sense; just a description of the study population

oucru

UNIVERSITY OF OXFORD

wellcome

When and how to use p-values

- P-value linked to hypothesis about population characteristic
  - Do not use p-value in baseline table
- Significance level dichotomy is not a gold standard for statistical inference
  - Draw conclusions based on several considerations

SENS, SPEC, PPV, NPV

- SENS: probability to test positive given disease
- SPEC: probability to test negative given disease free
- PPV: probability to have disease given positive test
- NPV: probability not to have disease given negative test

Positive predictive value

- Woman aged 40 diagnosed with breast cancer by mammography
  Rare in women aged 40: $P(\text{cancer}) = 1\%$
- Mammography is not perfect *($+$ or $-$: result from mammography)*
  - 20% false negative: $P(+|\text{cancer}) = 80\%$ (sensitivity)
  - 9.6% false positive: $P(-|\text{no cancer}) = 90.4\%$ (specificity)
- What is the probability that she has breast cancer?

Answer?

1. 0% - 30%?
2. 30% - 60%?
3. 60% - 100%?

Answer?

1. 0% - 30%?
2. 30% - 60%?
3. 60% - 100%?

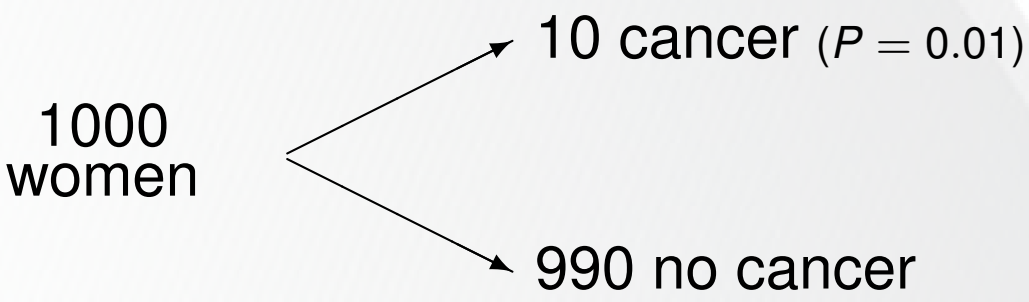
7.8%

95 out of 100 doctors said between 70% and 80% (reference unknown)

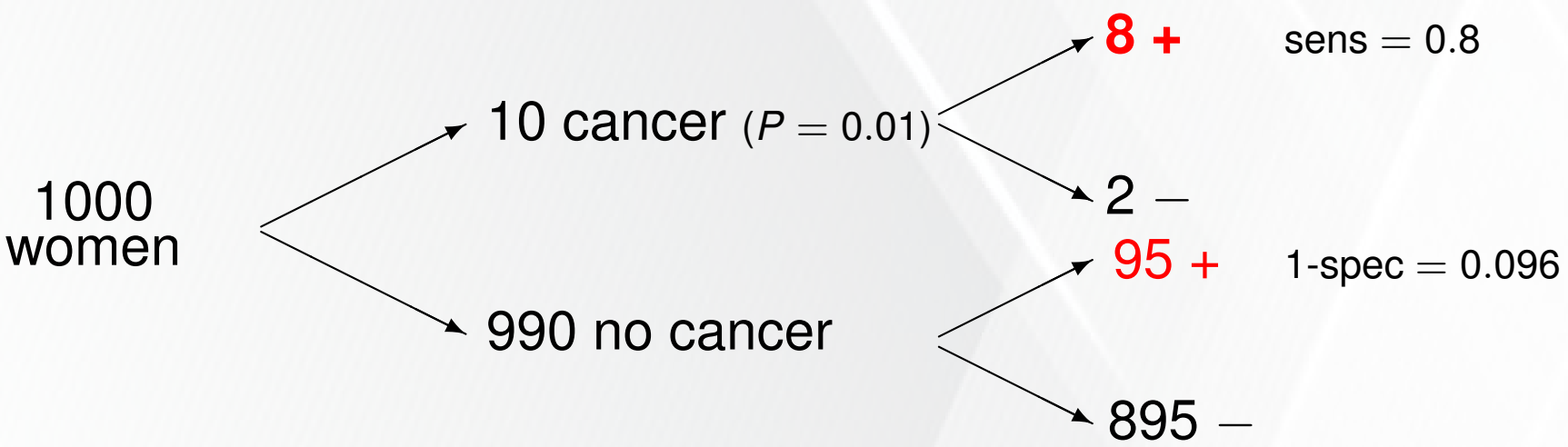
Positive predictive value

- Woman aged 40 diagnosed with breast cancer by mammography
  Rare in women aged 40: $P(\text{cancer}) = 1\%$
- Mammography is not perfect *($+$ or $-$: result from mammography)*
  - 20% false negative: $P(+|\text{cancer}) = 80\%$ (sensitivity)
  - 9.6% false positive: $P(-|\text{no cancer}) = 90.4\%$ (specificity)
- What is the probability that she has breast cancer?
  She needs to know $P(\text{cancer}|+)$

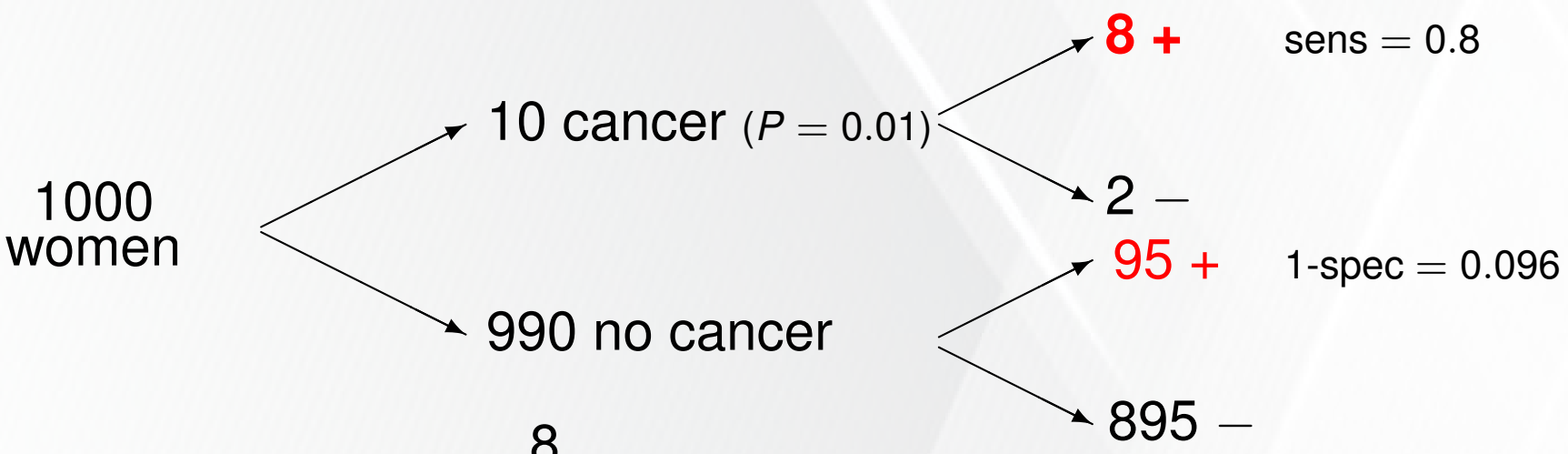
Answer, using large numbers



Answer, using large numbers



Answer, using large numbers



$$P(\text{cancer}|+) = \frac{8}{95+8} = 0.078$$

$8 + 95 = 103$ test positive;
8 of these have cancer
Positive predictive value (PPV) = 7.8%

What's wrong with significance tests (Sterne *et al.*, BMJ 2001)

http://www.bmj.com/content/322/7280/226.1

What's wrong with significance tests (Sterne *et al.*, BMJ 2001)

http://www.bmj.com/content/322/7280/226.1

Example: Suppose we perform 1000 comparable experiments.

Result in experiment	H_0 true (no effect)	H_0 false (effect)	Total
H_0 not rejected			
H_0 rejected			
Total			1000

What's wrong with significance tests (Sterne *et al.*, BMJ 2001)

http://www.bmj.com/content/322/7280/226.1

Example: Suppose we perform 1000 comparable experiments. Assume H_0 holds in 90% of them (no effect).

Result in experiment	H_0 true (no effect)	H_0 false (effect)	Total
H_0 not rejected			
H_0 rejected			
Total	900	100	1000

What's wrong with significance tests (Sterne *et al.*, BMJ 2001)

http://www.bmj.com/content/322/7280/226.1

Example: Suppose we perform 1000 comparable experiments.
Assume H_0 holds in 90% of them (no effect).
Assume level of significance 5%

Result in experiment	H_0 true (no effect)	H_0 false (effect)	Total
H_0 not rejected	855		
H_0 rejected	45		
Total	900	100	1000

oucru

UNIVERSITY OF OXFORD

wellcome

What's wrong with significance tests (Sterne *et al.*, BMJ 2001)

http://www.bmj.com/content/322/7280/226.1

Example: Suppose we perform 1000 comparable experiments.
Assume H_0 holds in 90% of them (no effect).
Assume level of significance 5% and power 50%

Result in experiment	H_0 true (no effect)	H_0 false (effect)	Total
H_0 not rejected	855	50	
H_0 rejected	45	50	
Total	900	100	1000

What's wrong with significance tests (Sterne *et al.*, BMJ 2001)

http://www.bmj.com/content/322/7280/226.1

Example: Suppose we perform 1000 comparable experiments.
Assume H_0 holds in 90% of them (no effect).
Assume level of significance 5% and power 50%

Result in experiment	H_0 true (no effect)	H_0 false (effect)	Total
H_0 not rejected	855	50	905
H_0 rejected	45	50	95
Total	900	100	1000

In only 50 out of 95, rejection of null hypothesis is correct!!

What's wrong with significance tests (Sterne *et al.*, BMJ 2001)

http://www.bmj.com/content/322/7280/226.1

Example: Suppose we perform 1000 comparable experiments.
Assume H_0 holds in 90% of them (no effect).
Assume level of significance 5% and power 50%

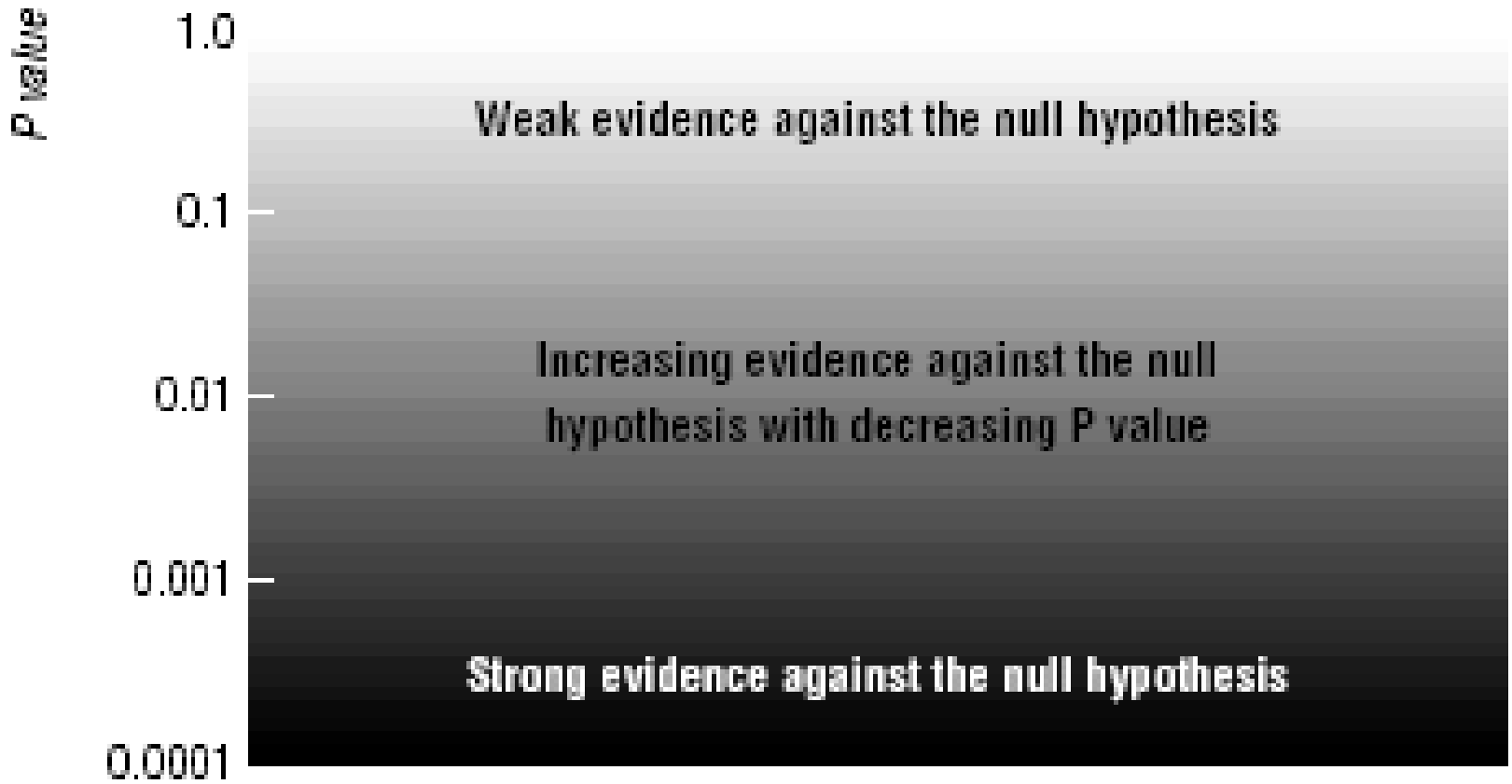
Result in experiment	H_0 true (no effect)	H_0 false (effect)	Total
H_0 not rejected	855	50	905
H_0 rejected	45	50	95
Total	900	100	1000

In only 50 out of 95, rejection of null hypothesis is correct!!
Do not test randomly, first think about biological plausibility
(statistical program is not a p-value generator)

Power of study (proportion (%) of time we reject null hypothesis if it is false)	Percentage of "significant" results that are false positives		
	P=0.05	P=0.01	P=0.001
80% of ideas correct (null hypothesis false)			
20	5.9	1.2	0.10
50	2.4	0.5	0.05
80	1.5	0.3	0.03
50% of ideas correct (null hypothesis false)			
20	20.0	4.8	0.50
50	9.1	2.0	0.20
80	5.9	1.2	0.10
10% of ideas correct (null hypothesis false)			
20	69.2	31.0	4.30
50	47.4*	15.3	1.80
80	36.0	10.1	1.10
1% of ideas correct (null hypothesis false)			
20	96.1	83.2	33.10
50	90.8	66.4	16.50
80	86.1	55.3	11.00

*Corresponds to assumptions in table 2.

$p = 0.05$ is to large extent an arbitrary choice



Suggested interpretation of P values from published medical research

Statement American Statistical Association

THE AMERICAN STATISTICIAN
2016, VOL. 70, NO. 2, 129–133
http://dx.doi.org/10.1080/00031305.2016.1154108

Taylor & Francis
Taylor & Francis Group

EDITORIAL

The ASA's Statement on *p*-Values: Context, Process, and Purpose

In February 2014, George Cobb, Professor Emeritus of Mathematics and Statistics at Mount Holyoke College, posed these questions to an ASA discussion forum:

> Q: Why do so many colleges and grad schools teach $p = 0.05$?
> A: Because that's still what the scientific community and journal editors use.
> Q: Why do so many people still use $p = 0.05$?
> A: Because that's what they were taught in college or grad school.

Cobb's concern was a long-worrisome circularity in the sociology of science based on the use of bright lines such as $p < 0.05$: "We teach it because it's what we do; we do it because it's what we teach." This concern was brought to the attention of the ASA Board.

2014) and a statement on risk-limiting post-election audits (American Statistical Association 2010). However, these were truly policy-related statements. The VAM statement addressed a key educational policy issue, acknowledging the complexity of the issues involved, citing limitations of VAMs as effective performance models, and urging that they be developed and interpreted with the involvement of statisticians. The statement on election auditing was also in response to a major but specific policy issue (close elections in 2008), and said that statistically based election audits should become a routine part of election processes.

By contrast, the Board envisioned that the ASA statement on *p*-values and statistical significance would shed light on an

Some ASA statements

- The widespread use of “statistical significance” (generally interpreted as $p \leq 0.05$) as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process.

Some ASA statements

- The widespread use of “statistical significance” (generally interpreted as $p \leq 0.05$) as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process.
- Researchers should disclose the number of hypotheses explored during the study, all data collection decisions, all statistical analyses conducted, and all p-values computed.

Some ASA statements

- The widespread use of "statistical significance" (generally interpreted as $p \leq 0.05$) as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process.
- Researchers should disclose the number of hypotheses explored during the study, all data collection decisions, all statistical analyses conducted, and all p-values computed.
- Good statistical practice, as an essential component of good scientific practice, emphasizes principles of good study design and conduct, a variety of numerical and graphical summaries of data, understanding of the phenomenon under study, interpretation of results in context, complete reporting and proper logical and quantitative understanding of what data summaries mean. No single index should substitute for scientific reasoning.

Outline

Which variables to include in our regression model
- Variable reduction
- Multiple testing and fishing expeditions

How to include variables
- Stratified analysis
- Dichotomania

The role of p-values
- Examples

Reporting

Help!

oucru

UNIVERSITY OF OXFORD

wellcome

Criteria for reviewers (PLOS neglected tropical diseases)

Results:

- Does the analysis presented match the analysis plan?
- Are the results clearly and completely presented?
- Are the figures (Tables, Images) of sufficient quality for clarity?

Conclusions:

- Are the conclusions supported by the data presented?
- Are the limitations of analysis clearly described?
- Do the authors discuss how these data can be helpful to advance our understanding of the topic under study?
- Is public health relevance addressed?

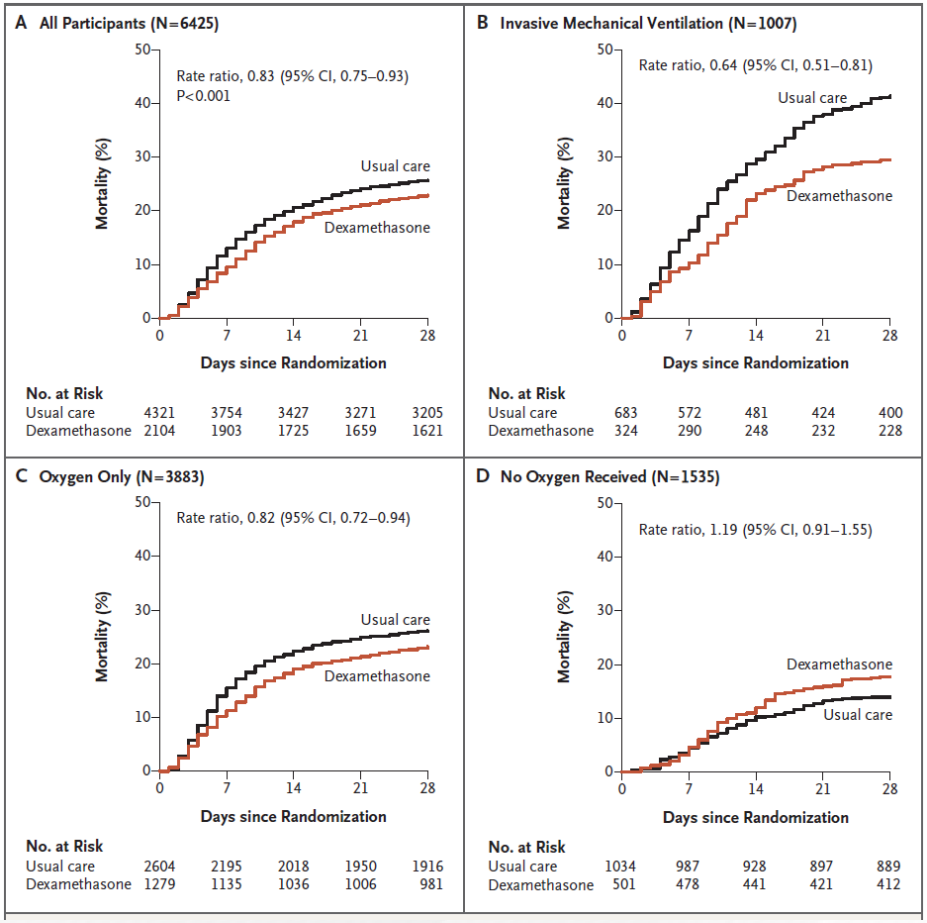
Risk factors for anal carcinoma caused by HPV

- HPV: human papilloma virus
- Give p-value instead of “NS” (not significant)
- Also, we don’t want to quantify how age differs by outcome, we want to know how age influences outcome (reverse causal order)

Table 1. Characteristics of patients according to high-grade dysplasia and cancer outcomes.

	With high-grade dysplasia and cancer		
	Yes	No	*P*
n	38	161	
At baseline			
Age (years) (mean ± SD)	37 ± 10	36.9 ± 11	NS
Male (%)	94	82	NS
CD4 T-cell count ($\times 10^6$ cells/ml serum) (mean ± SD)	362 ± 319	455 ± 340	NS
Langerhans’ cells/mm^2 mucosa (mean ± SD)	16 ± 13	27.4 ± 24	0.01
HIV positive (%)	84	16	0.007
Oncogenic human papillomavirus subtype (%)	33	1.6	0.009
Epstein–Barr virus (%)	12	3.3	0.05
Herpes simplex virus (%)	39	12.5	0.007
Anal co-infection (%)	51	18	0.0002
During follow up			
CD4 T-cell count ($\times 10^6$ cells/ml serum)[a] (mean ± SD)	334 ± 270	360 ± 312	NS
Anal infections (mean ± SD)	1.8 ± 0.7	1.3 ± 0.8	< 0.01
Relapses (mean ± SD)	2.4 ± 0.6	1.9 ± 0.8	0.05
Langerhans’ cells/mm^2 mucosa[a,b] (mean ± SD)	7 ± 6	19 ± 9	0.001
Oncogenic human papillomavirus subtype (%)	29	2.2	< 0.001

Recovery (dexamethasone in Covid-19) trial



In patients hospitalized with Covid-19, the use of dexamethasone resulted in lower 28-day mortality among those who were receiving either invasive mechanical ventilation or oxygen alone at randomization but not among those receiving no respiratory support.

SXT versus azithromycin in patients with undifferentiated febrile illness

- *We hypothesized that azithromycin is superior to SXT for UFI treatment, but the drugs are non-inferior to each other for culture-positive enteric fever treatment*
  culture-positive: Salmonella Typhi or paratyphoid fever

SXT versus azithromycin in patients with undifferentiated febrile illness

- *We hypothesized that azithromycin is superior to SXT for UFI treatment, but the drugs are non-inferior to each other for culture-positive enteric fever treatment*
  culture-positive: Salmonella Typhi or paratyphoid fever
- What is the null hypothesis?
  What is the population?

SXT versus azithromycin in patients with undifferentiated febrile illness

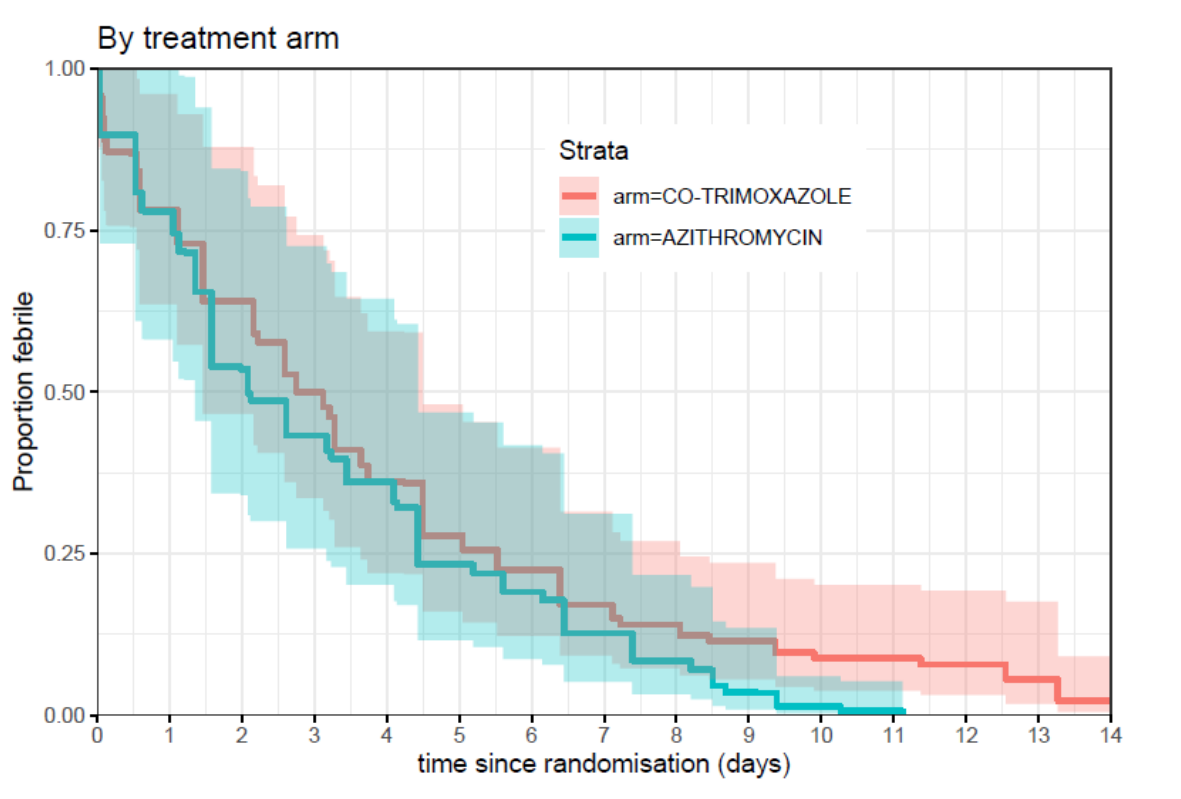
- *We hypothesized that azithromycin is superior to SXT for UFI treatment, but the drugs are non-inferior to each other for culture-positive enteric fever treatment*
  culture-positive: Salmonella Typhi or paratyphoid fever
- What is the null hypothesis?
  What is the population?
- There are two hypotheses, one for overall population of patients with UFI and the other for the culture-confirmed subgroup

SXT versus azithromycin in patients with undifferentiated febrile illness

- *We hypothesized that azithromycin is superior to SXT for UFI treatment, but the drugs are non-inferior to each other for culture-positive enteric fever treatment*
  culture-positive: Salmonella Typhi or paratyphoid fever
- What is the null hypothesis?
  What is the population?
- There are two hypotheses, one for overall population of patients with UFI and the other for the culture-confirmed subgroup
- If superior for overall population, but non-inferior for culture-positive subgroup, then it must be superior for the culture-negative subgroup

SXT versus azithromycin in patients with UFI: results

Despite similar fever clearance time in the two arms (primary outcome, P-value: 0.059), significantly fewer complications and relapses make azithromycin a better choice for empirical treatment of UFI in Nepal



oucru

UNIVERSITY OF OXFORD

wellcome

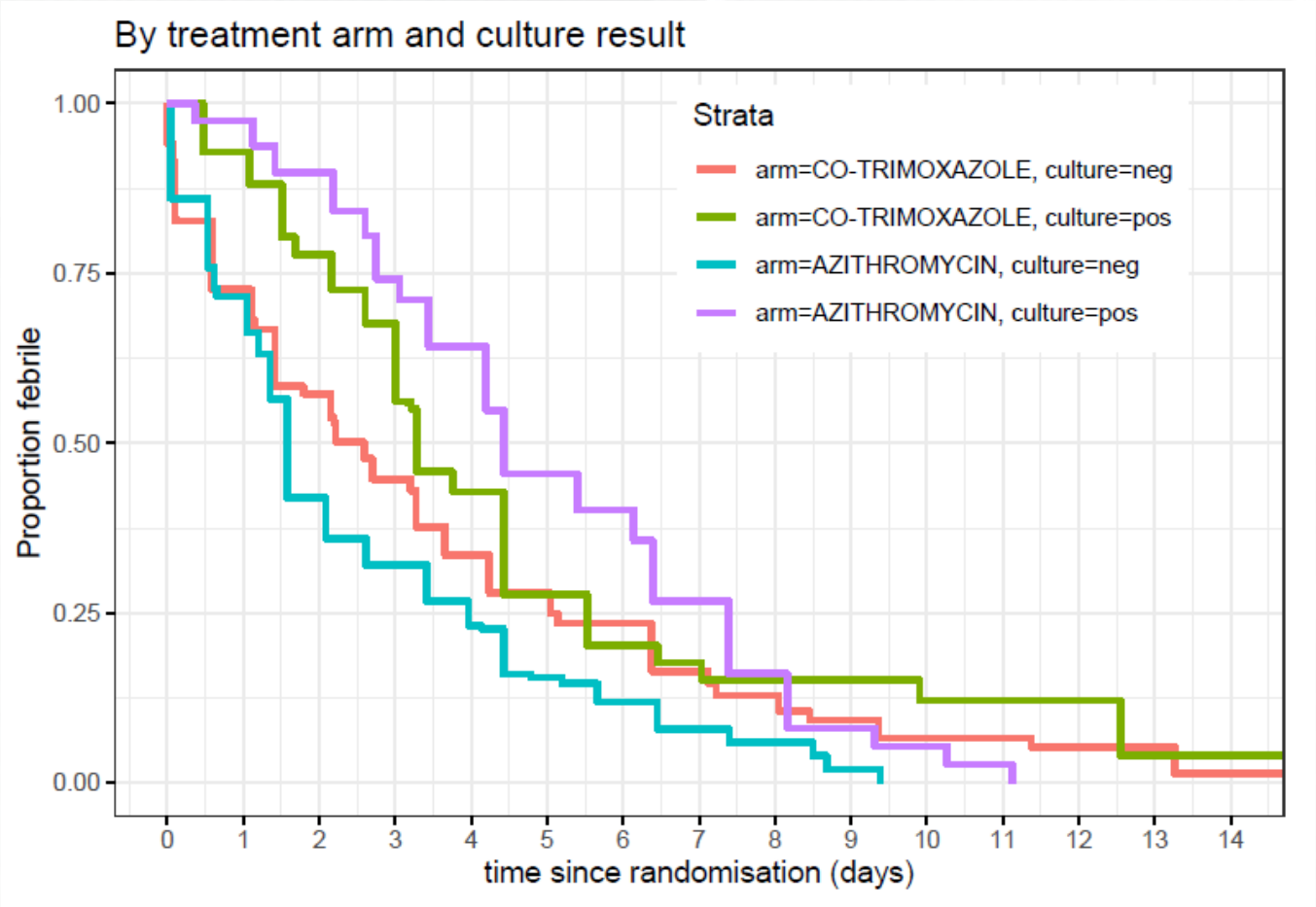
My suggestion

We found moderate evidence that seven days of azithromycin is more effective than 7 days co-trimoxazole for the treatment of all cause UFI in Nepal.

Comment by reviewer: *The claims about differences in fever clearance times for all-cause fever in the SXT group compared with the azithromycin group are overstated. These differences were not statistically significant. The language should be revised for a more honest rendering of the main findings.*

By culture

P-value: 0.024 in culture negative; 0.81 in culture positive



Outline

Which variables to include in our regression model
- Variable reduction
- Multiple testing and fishing expeditions

How to include variables
- Stratified analysis
- Dichotomania

The role of p-values
- Examples

Reporting

Help!

Reporting

- Report units for measurements (also in axis labels)
- Report N's (denominators) and the amount of missing data
- For effect estimates, report estimates and CI in addition to p-values
- Use reasonable precision for reporting, no "fake precision"
  - Usually, 2 decimals for p-values are enough, except if they are very small (e.g. $p \leq 0.0001$)
  - For %: usually, full numbers or max 2 decimals are enough
    - Good example: "17 (74%) of 23 patients had a clinical response"
    - Bad example: "17 (73.913%) of 23 patients. . ."

Criteria for reviewers (PLOS neglected tropical diseases)

Results:

- Does the analysis presented match the analysis plan?
- Are the results clearly and completely presented?
- Are the figures (Tables, Images) of sufficient quality for clarity?

Conclusions:

- Are the conclusions supported by the data presented?
- Are the limitations of analysis clearly described?
- Do the authors discuss how these data can be helpful to advance our understanding of the topic under study?
- Is public health relevance addressed?

Criteria for reviewers (PLOS neglected tropical diseases)

Results:

- Does the analysis presented match the analysis plan?
- Are the results clearly and completely presented?
- Are the figures (Tables, Images) of sufficient quality for clarity?

Conclusions:

- Are the conclusions supported by the data presented?
- Are the limitations of analysis clearly described?
- Do the authors discuss how these data can be helpful to advance our understanding of the topic under study?
- Is public health relevance addressed?

Think about rescaling of numeric variable

- Parameter is measure per unit increase
- Choose informative unit. Not as in e.g.

  Incidence and clearance of genital HPV infection in men

	Any HPV		Oncogenic HPV		Non-oncogenic HPV	
	Univariate	Multivariate*	Univariate	Multivariate*	Univariate	Multivariate†
Country						
USA	1·00	1·00	1·00	1·00	1·00	1·00
Brazil	1·07 (0·82–1·40)	0·93 (0·69–1·27)	0·89 (0·69–1·15)	0·81 (0·61–1·09)	1·56 (1·2–2·03)	2·04 (1·47–2·83)
Mexico	0·82 (0·63–1·08)	0·83 (0·63–1·10)	0·65 (0·49–0·86)	0·75 (0·56–1·00)	0·99 (0·75–1·31)	1·27 (0·90–1·78)
Age	1·00 (0·99–1·00)	0·99 (0·98–1·00)	0·99 (0·98–1·00)	0·99 (0·98–1·00)	1·00 (0·99–1·01)	0·99 (0·98–1·00)

- Better show age effect per ten years (or use more digits)

oucru

UNIVERSITY OF OXFORD

wellcome

Reporting (continued)

- Follow standard reporting guidelines
- Consort for RCTs (www.consort-statement.org)
- Strobe for observational studies (www.strobe-statement.org)
- STARD for diagnostic tests (www.stard-statement.org)
- More on https://www.equator-network.org

Outline

Which variables to include in our regression model
- Variable reduction
- Multiple testing and fishing expeditions

How to include variables
- Stratified analysis
- Dichotomania

The role of p-values
- Examples

Reporting

Help!

Analysis and reporting

- Reserve enough time for a careful analysis
- Data quality checks, missing data, outliers
- Use descriptive and graphical analyses to understand the data
- Structured analysis strategy (following the analysis plan)
- Restrict fishing expeditions; be aware that they produce at best exploratory/preliminary evidence
- Report descriptive statistics first, then the primary analysis, then secondary analyses
- Report interesting exploratory analyses but clearly declare them as hypothesis-generating only

When do you need a statistician?

- Involve a statistician as early as possible, i.e. during the design stage of the trial (not only at the analysis stage)!
- Involve a statistician for all studies that are large, have a complex design or will require a substantial amount of statistical analyses
- Complex data structures or analyses
- Longitudinal data, survival data, clustered data
- High-dimensional data
- Large amount of missing data
- Whenever you feel insecure about the correct design or analysis
- Prevent miscommunication

Recommended books (entire course)

- Altman DG (1991). *Practical Statistics for Medical Research* Chapman & Hall/CRC
- Neale Batra *The Epidemiologist R Handbook*. https://epirhandbook.com/en/ Also in Vietnamese!
- Frank E. Harrell (2022). *Biostatistics for Biomedical Research* http://hbiostat.org/bbr/
- Katz MH (2006). *Study Design and Statistical Analysis* Cambridge University Press
- Kirkwood BR and Sterne JAC (2003). *Essential Medical Statistics* (2nd Edition). Blackwell Science.
- Vu J and Harrington D (2020). *Introductory Statistics for the Life and Biomedical Sciences* OpenIntro https://www.openintro.org/book/biostat/