p-hacking strategies

Presenter: Bui Minh Tri

Objective + Outline

- Objective

p-hacking strategies

- Outline

1. **Selective reporting DV**

2. **Selective reporting IV**

Introduction

p-hacking

- Compound strategies: non-significant ⇨ significant result.

- Not every researcher aware of it.
  - ➢ **Not necessarily an intentional attempt at gaming the system.**

1. Selective reporting dependent variable

Background

- **p-hacking**:
  - ✓ Treatment vs control group: compare different outcome/ dependent variables.

Background

- **p-hacking**:
  - ✓ Treatment vs control group: compare different outcome/ dependent variables.

Independent

Table 1b | Efficacy of surgical face masks in reducing respiratory virus frequency of detection and viral shedding in respiratory droplets and aerosols of symptomatic individuals with coronavirus, influenza virus or rhinovirus infection

	Droplet particles >5 µm			Aerosol particles ≤5 µm		
Virus type	**Without surgical face mask**	**With surgical face mask**	***P***	**Without surgical face mask**	**With surgical face mask**	***P***
	Detection of virus					
	No. positive/no. total (%)	**No. positive/no. total (%)**		**No. positive/no. total (%)**	**No. positive/no. total (%)**	
Coronavirus	3 of 10 (30)	0 of 11 (0)	0.09	**4 of 10 (40)**	**0 of 11 (0)**	**0.04**
Influenza virus	**6 of 23 (26)**	**1 of 27 (4)**	**0.04**	8 of 23 (35)	6 of 27 (22)	0.36
Rhinovirus	9 of 32 (28)	6 of 27 (22)	0.77	19 of 34 (56)	12 of 32 (38)	0.15
	Viral load ($\log_{10}$ virus copies per sample)					
	Median (IQR)	**Median (IQR)**		**Median (IQR)**	**Median (IQR)**	
Coronavirus	0.3 (0.3, 1.2)	0.3 (0.3, 0.3)	0.07	**0.3 (0.3, 3.3)**	**0.3 (0.3, 0.3)**	**0.02**
Influenza virus	**0.3 (0.3, 1.1)**	**0.3 (0.3, 0.3)**	**0.01**	0.3 (0.3, 3.0)	0.3 (0.3, 0.3)	0.26
Rhinovirus	0.3 (0.3, 1.3)	0.3 (0.3, 0.3)	0.44	1.8 (0.3, 2.8)	0.3 (0.3, 2.4)	0.12

P values for comparing the frequency of respiratory virus detection between the mask intervention were obtained by two-sided Fisher's exact test and (two-sided) *P* values for mask intervention as predictor of $\log_{10}$ virus copies per sample were obtained by an unadjusted univariate Tobit regression model, which allowed for censoring at the lower limit of detection of the RT-PCR assay, with significant differences in bold. Undetectable values were imputed as 0.3 $\log_{10}$ virus copies per sample. IQR, interquartile range.

Leung NH, Chu DK, Shiu EY, Chan KH, McDevitt JJ, Hau BJ, Yen HL, Li Y, Ip DK, Peiris JS, Seto WH. Respiratory virus shedding in exhaled breath and efficacy of face masks. Nature medicine. 2020 May;26(5):676-80.

Background

- **p-hacking**:
  - ✓ Treatment vs control group: compare different outcome/ dependent variables.

Independent

Dependent

Table 1b | Efficacy of surgical face masks in reducing respiratory virus frequency of detection and viral shedding in respiratory droplets and aerosols of symptomatic individuals with coronavirus, influenza virus or rhinovirus infection

	Droplet particles >5 µm			Aerosol particles ≤5 µm		
Virus type	**Without surgical face mask**	**With surgical face mask**	***P***	**Without surgical face mask**	**With surgical face mask**	***P***
	Detection of virus					
	No. positive/no. total (%)	**No. positive/no. total (%)**		**No. positive/no. total (%)**	**No. positive/no. total (%)**	
Coronavirus	3 of 10 (30)	0 of 11 (0)	0.09	**4 of 10 (40)**	**0 of 11 (0)**	**0.04**
Influenza virus	**6 of 23 (26)**	**1 of 27 (4)**	**0.04**	8 of 23 (35)	6 of 27 (22)	0.36
Rhinovirus	9 of 32 (28)	6 of 27 (22)	0.77	19 of 34 (56)	12 of 32 (38)	0.15
	Viral load ($\log_{10}$ virus copies per sample)					
	Median (IQR)	**Median (IQR)**		**Median (IQR)**	**Median (IQR)**	
Coronavirus	0.3 (0.3, 1.2)	0.3 (0.3, 0.3)	0.07	**0.3 (0.3, 3.3)**	**0.3 (0.3, 0.3)**	**0.02**
Influenza virus	**0.3 (0.3, 1.1)**	**0.3 (0.3, 0.3)**	**0.01**	0.3 (0.3, 3.0)	0.3 (0.3, 0.3)	0.26
Rhinovirus	0.3 (0.3, 1.3)	0.3 (0.3, 0.3)	0.44	1.8 (0.3, 2.8)	0.3 (0.3, 2.4)	0.12

P values for comparing the frequency of respiratory virus detection between the mask intervention were obtained by two-sided Fisher's exact test and (two-sided) *P* values for mask intervention as predictor of $\log_{10}$ virus copies per sample were obtained by an unadjusted univariate Tobit regression model, which allowed for censoring at the lower limit of detection of the RT-PCR assay, with significant differences in bold. Undetectable values were imputed as 0.3 $\log_{10}$ virus copies per sample. IQR, interquartile range.

Leung NH, Chu DK, Shiu EY, Chan KH, McDevitt JJ, Hau BJ, Yen HL, Li Y, Ip DK, Peiris JS, Seto WH. Respiratory virus shedding in exhaled breath and efficacy of face masks. Nature medicine. 2020 May;26(5):676-80.

Background

- **p-hacking**:
  - ✓ Treatment vs control group: compare different outcome/ dependent variables.
  - ✓ Conduct 1 hypothesis test for each dependent variable.

Table 1b | Efficacy of surgical face masks in reducing respiratory virus frequency of detection and viral shedding in respiratory droplets and aerosols of symptomatic individuals with coronavirus, influenza virus or rhinovirus infection

	Droplet particles >5 μm				Aerosol particles ≤5 μm			
Virus type	Without surgical face mask	With surgical face mask		*P*	Without surgical face mask	With surgical face mask	*P*	
	Detection of virus							
	No. positive/no. total (%)	No. positive/no. total (%)			No. positive/no. total (%)	No. positive/no. total (%)		
Coronavirus	3 of 10 (30)	0 of 11 (0)	*1*	0.09	**4 of 10 (40)**	**0 of 11 (0)**	**0.04**	*7*
Influenza virus	**6 of 23 (26)**	**1 of 27 (4)**	*2*	**0.04**	8 of 23 (35)	6 of 27 (22)	0.36	*8*
Rhinovirus	9 of 32 (28)	6 of 27 (22)	*3*	0.77	19 of 34 (56)	12 of 32 (38)	0.15	*9*
	Viral load ($\log_{10}$ virus copies per sample)							
	Median (IQR)	Median (IQR)			Median (IQR)	Median (IQR)		
Coronavirus	0.3 (0.3, 1.2)	0.3 (0.3, 0.3)	*4*	0.07	**0.3 (0.3, 3.3)**	**0.3 (0.3, 0.3)**	**0.02**	*10*
Influenza virus	**0.3 (0.3, 1.1)**	**0.3 (0.3, 0.3)**	*5*	**0.01**	0.3 (0.3, 3.0)	0.3 (0.3, 0.3)	0.26	*11*
Rhinovirus	0.3 (0.3, 1.3)	0.3 (0.3, 0.3)	*6*	0.44	1.8 (0.3, 2.8)	0.3 (0.3, 2.4)	0.12	*12*

P values for comparing the frequency of respiratory virus detection between the mask intervention were obtained by two-sided Fisher's exact test and (two-sided) *P* values for mask intervention as predictor of $\log_{10}$ virus copies per sample were obtained by an unadjusted univariate Tobit regression model, which allowed for censoring at the lower limit of detection of the RT-PCR assay, with significant differences in bold. Undetectable values were imputed as 0.3 $\log_{10}$ virus copies per sample. IQR, interquartile range.

Background

- **p-hacking**:
  - ✓ Treatment vs control group: compare different outcome/ dependent variables.
  - ✓ Conduct 1 hypothesis test for each dependent variable.
  - ✓ Selectively report the significant results.

1. Selective reporting the dependent variable

- Assume using t-test.
- **FPR from 3 – 10 dependent variables ?**

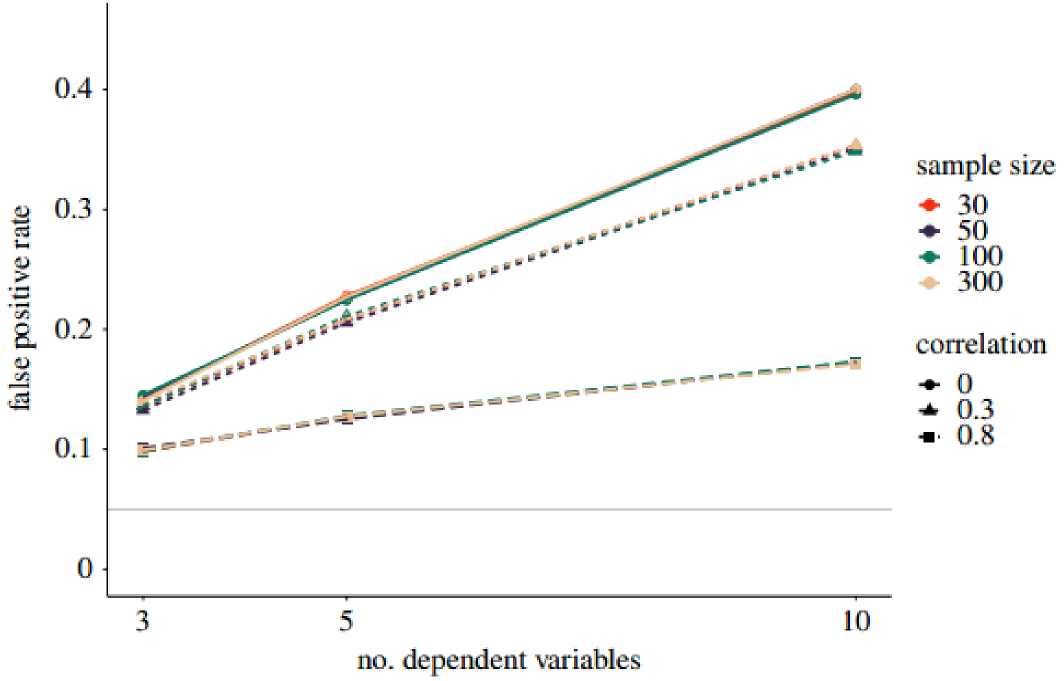
1. Selective reporting the dependent variable

- Assume using t-test.
- **FPR from 3 – 10 dependent variables ?**

	Virus Type	Without mask	With mask	p
1	Coronavirus	0.3 (0.3, 1.2)	0.3 (0.3, 0.3)	0.07
2	Influenza virus	0.3 (0.3, 1.1)	0.3 (0.3, 0.3)	0.01
…	….	….	….	….
10	Rhino virus	0.3 (0.3, 1.3)	0.3 (0.3, 0.3)	0.44

1. Selective reporting the dependent variable

- Assume using t-test.
- **FPR from 3 – 10 dependent variables ?**

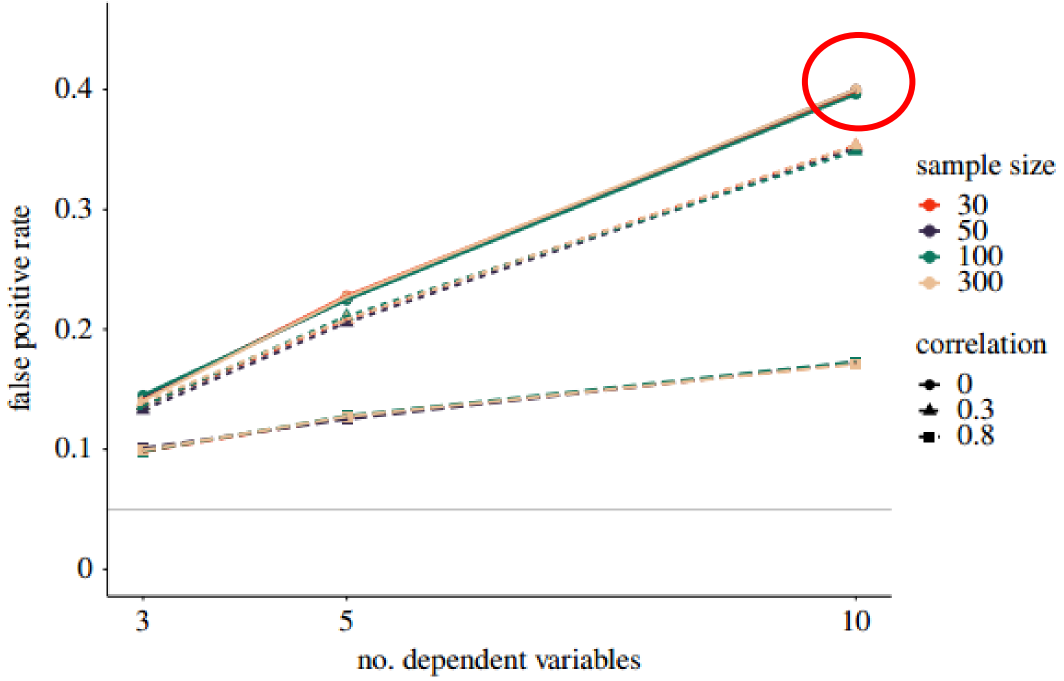


Number of dependent variables indicates how many hypothesis tests were conducted (at maximum) to obtain a significant result.

The solid grey line: nominal α-level of 5%.

1. Selective reporting the dependent variable

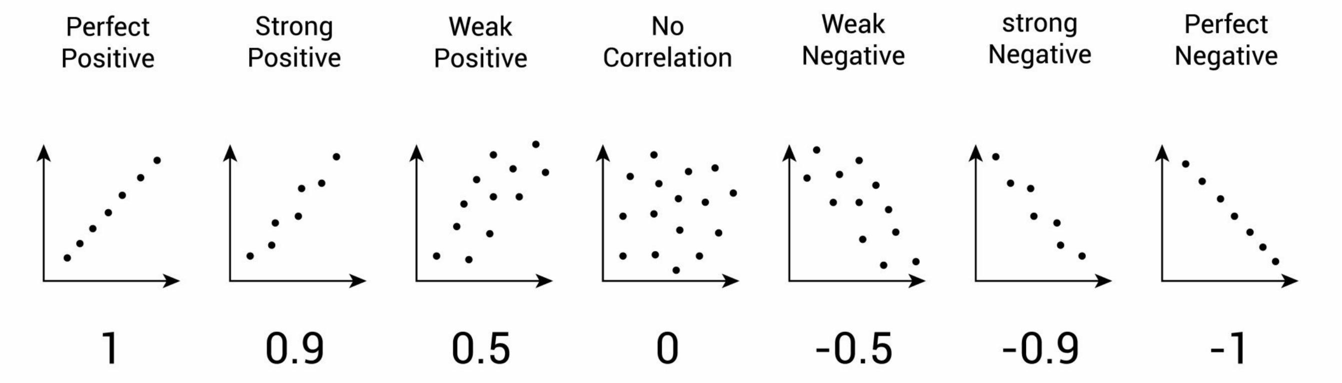
- Assume using t-test.
- Sample size: not a protective factor.
- 10 variables – **correlation = 0**: FPR ≈ 40%

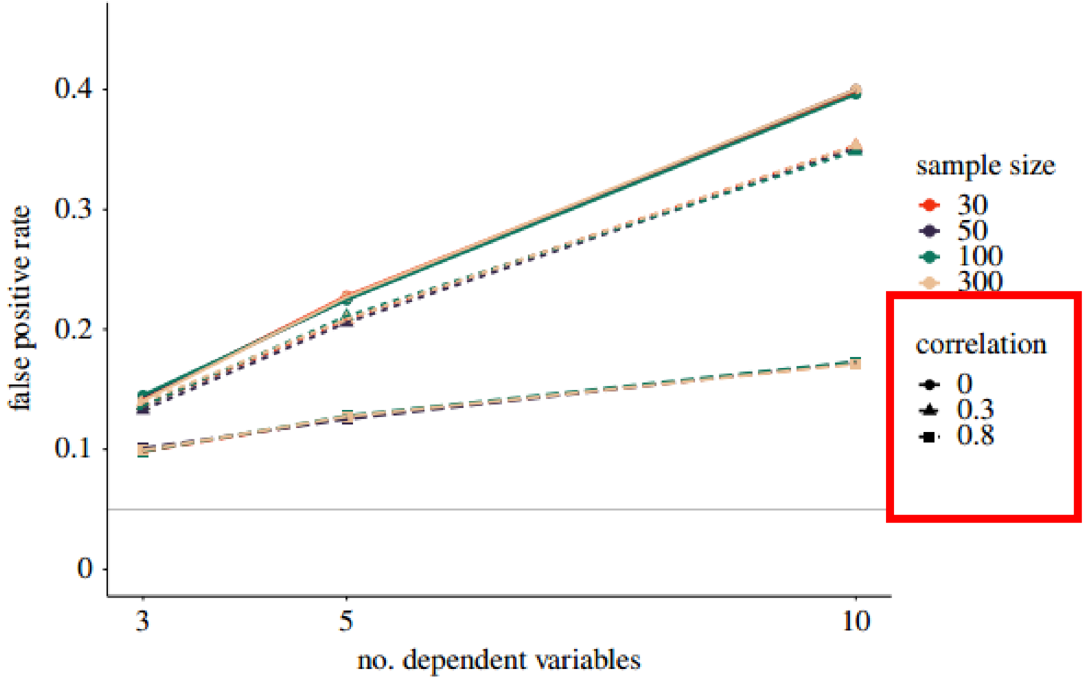


Number of dependent variables indicates how many hypothesis tests were conducted (at maximum) to obtain a significant result.
The solid grey line: nominal α-level of 5%.

What is correlation ?

- Relationship **2 quantitative variables**
- **Correlation coefficient (r)**





Number of dependent variables indicates how many hypothesis tests were conducted (at maximum) to obtain a significant result.
The solid grey line: nominal α-level of 5%.

What is correlation ?

	Case N=503	Control N=493	p value
Height	158 [153;165]	158 [154;165]	0.662
Weight	62.0 [55.0;70.0]	58.0 [52.8;65.0]	<0.001
BMI	24.3 [22.4;27.2]	23.2 [21.1;25.4]	<0.001
Waist	86.0 [80.0;93.0]	82.0 [75.0;88.0]	<0.001
Hip	95.0 [89.0;100]	92.0 [86.0;97.0]	<0.001

What is correlation ?

	Case N=503	Control N=493	p value
Height	158 [153;165]	158 [154;165]	0.662
Weight	62.0 [55.0;70.0]	58.0 [52.8;65.0]	<0.001
BMI	24.3 [22.4;27.2]	23.2 [21.1;25.4]	<0.001
Waist	86.0 [80.0;93.0]	82.0 [75.0;88.0]	<0.001
Hip	95.0 [89.0;100]	92.0 [86.0;97.0]	<0.001

Correlation Test

Correlation coefficient (r)

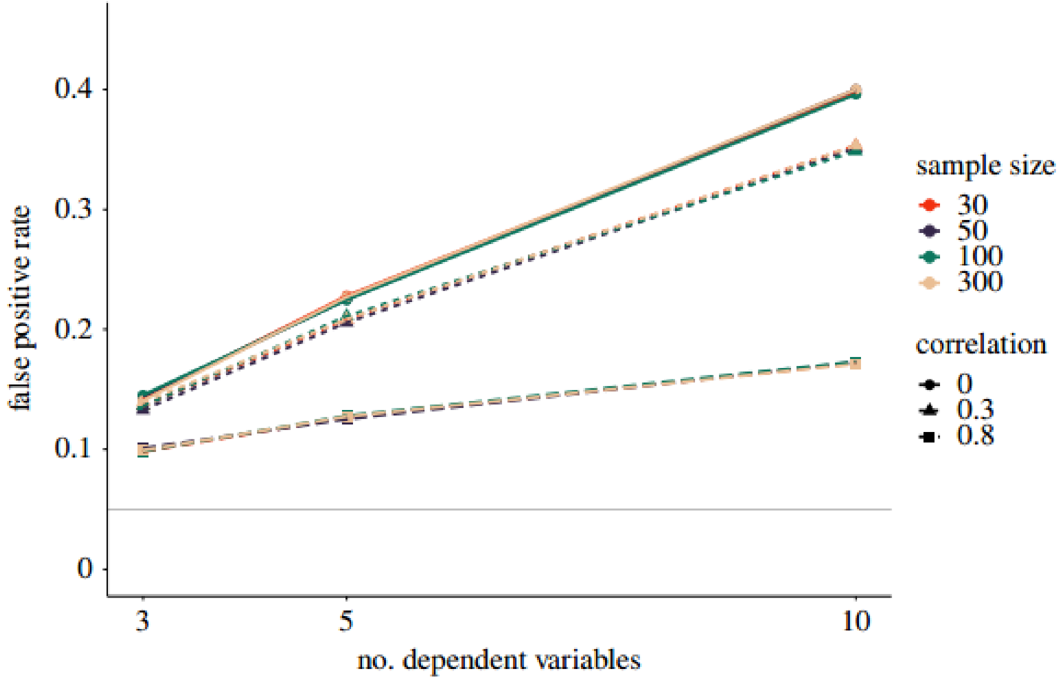
	height	weight	BMI	waist	hip
height	1.00	0.51	-0.02	0.20	0.19
weight	0.51	1.00	0.82	0.76	0.70
BMI	-0.02	0.82	1.00	0.77	0.70
waist	0.20	0.76	0.77	1.00	0.77
hip	0.19	0.70	0.70	0.77	1.00

P

	height	weight	BMI	waist	hip
height		0.0000	0.5275	0.0000	0.0000
weight	0.0000		0.0000	0.0000	0.0000
BMI	0.5275	0.0000		0.0000	0.0000
waist	0.0000	0.0000	0.0000		0.0000
hip	0.0000	0.0000	0.0000	0.0000	

1. Selective reporting the dependent variable

- Assume using t-test.
- Sample size: not a protective factor.
- 10 variables – **correlation = 0**: FPR ≈ 40%
- **FPR decreases** with:
  - ✓ **Less dependent variables**.
  - ✓ **Higher correlations variables**



Number of dependent variables indicates how many hypothesis tests were conducted (at maximum) to obtain a significant result.
The solid grey line: nominal α-level of 5%.

Multiple testing issue

- **Question**: ↑ hypothesis tests ⇨ ↑ False Positive Rate ?

Multiple testing issue

- If perform **m hypothesis independent tests**, **the probability at least 1 false positive** ?
  - ✓ P (Making Type I error) = α
  - ✓ P (Not making Type I error) = $1 - \alpha$
  - ✓ P (Not making an error in m tests) = $(1 - \alpha)^m$
  - ✓ P (Making at least 1 error in m tests) = $1 - (1 - \alpha)^m$

- Example: m = 100 tests, $\alpha = 0.05$ ⇨ $P = 1 - (1 - 0.05)^{100} = 0.99$
  - ➢ If have 100 hypothesis tests, the probability at least 1 false positive: 99%

Multiple testing issue

Table 1b | Efficacy of surgical face masks in reducing respiratory virus frequency of detection and viral shedding in respiratory droplets and aerosols of symptomatic individuals with coronavirus, influenza virus or rhinovirus infection

	Droplet particles >5 µm				Aerosol particles ≤5 µm			
Virus type	**Without surgical face mask**	**With surgical face mask**		***P***	**Without surgical face mask**	**With surgical face mask**		***P***
	Detection of virus							
	No. positive/no. total (%)	**No. positive/no. total (%)**			**No. positive/no. total (%)**	**No. positive/no. total (%)**		
Coronavirus	3 of 10 (30)	0 of 11 (0)	1	0.09	**4 of 10 (40)**	**0 of 11 (0)**	7	**0.04**
Influenza virus	**6 of 23 (26)**	**1 of 27 (4)**	2	**0.04**	8 of 23 (35)	6 of 27 (22)	8	0.36
Rhinovirus	9 of 32 (28)	6 of 27 (22)	3	0.77	19 of 34 (56)	12 of 32 (38)	9	0.15
	Viral load ($\log_{10}$ virus copies per sample)							
	Median (IQR)	**Median (IQR)**			**Median (IQR)**	**Median (IQR)**		
Coronavirus	0.3 (0.3, 1.2)	0.3 (0.3, 0.3)	4	0.07	**0.3 (0.3, 3.3)**	**0.3 (0.3, 0.3)**	10	**0.02**
Influenza virus	**0.3 (0.3, 1.1)**	**0.3 (0.3, 0.3)**	5	**0.01**	0.3 (0.3, 3.0)	0.3 (0.3, 0.3)	11	0.26
Rhinovirus	0.3 (0.3, 1.3)	0.3 (0.3, 0.3)	6	0.44	1.8 (0.3, 2.8)	0.3 (0.3, 2.4)	12	0.12

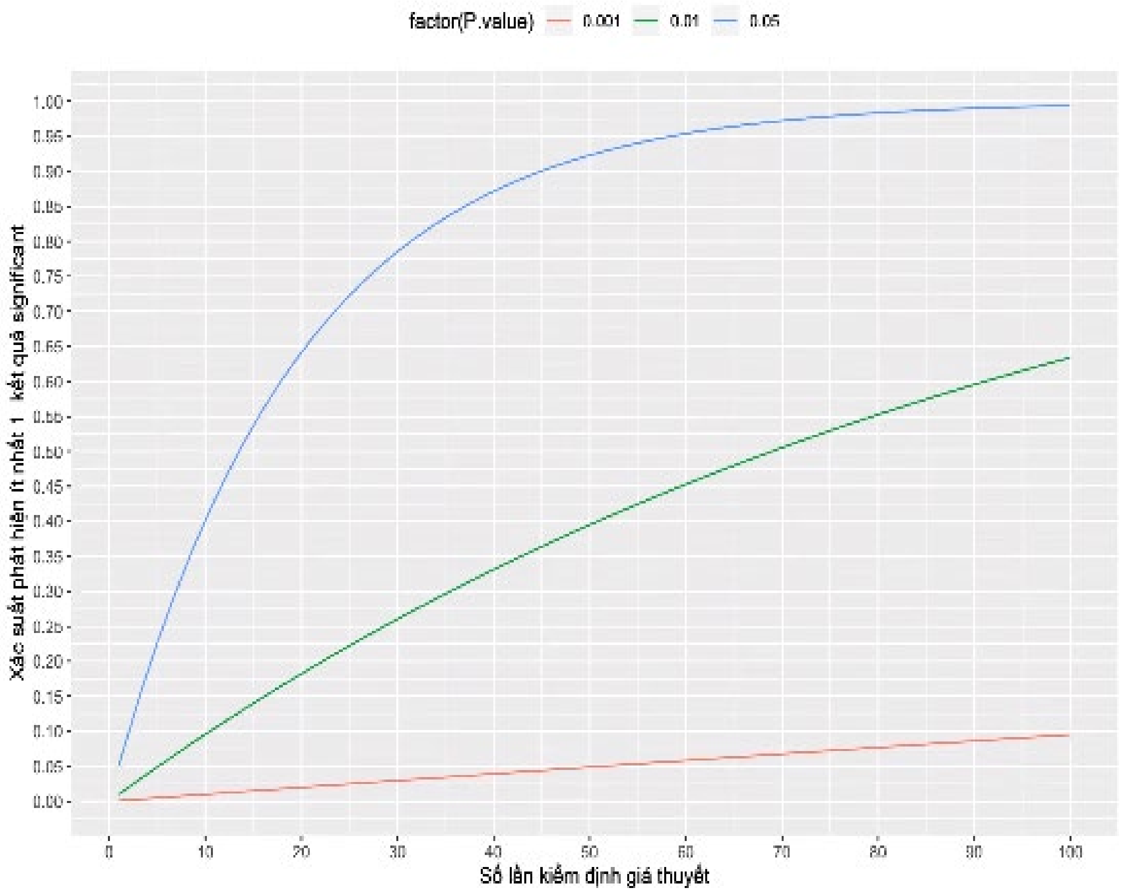
P values for comparing the frequency of respiratory virus detection between the mask intervention were obtained by two-sided Fisher's exact test and (two-sided) *P* values for mask intervention as predictor of $\log_{10}$ virus copies per sample were obtained by an unadjusted univariate Tobit regression model, which allowed for censoring at the lower limit of detection of the RT-PCR assay, with significant differences in bold. Undetectable values were imputed as 0.3 $\log_{10}$ virus copies per sample. IQR, interquartile range.

P (Making at least 1 error in m tests) $= 1 - (1 - \alpha)^m$

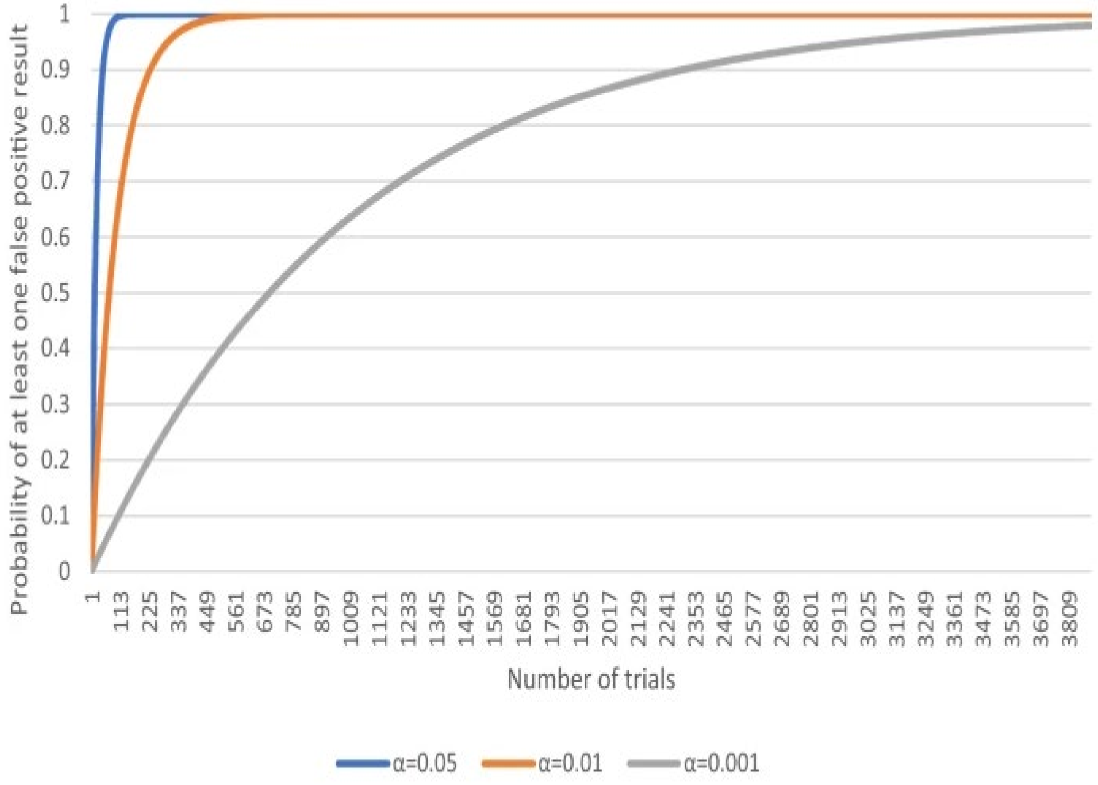
$= 1 - (1 - 0.05)^{12} = 0.4596 = 45.96\%$

Leung NH, Chu DK, Shiu EY, Chan KH, McDevitt JJ, Hau BJ, Yen HL, Li Y, Ip DK, Peiris JS, Seto WH. Respiratory virus shedding in exhaled breath and efficacy of face masks. Nature medicine. 2020 May;26(5):676-80.

Multiple testing issue



NguyenVanTuan-
https://www.youtube.com/watch?v=RPjVPHpeu2o&t=2517s



The probability of obtaining at least one false positive result $P(FP \geq 1)$ (own calculation)

Maziarz M, Stencel A. The failure of drug repurposing for COVID-19 as an effect of excessive hypothesis testing and weak mechanistic evidence. History and Philosophy of the Life Sciences. 2022 Dec;44(4):47.

Bonferroni Correction

- Bonferroni correction: $\alpha^* = \alpha / m$
  - ✓ α : significance level.
  - ✓ m : number of hypothesis tests.

Bonferroni Correction

- Bonferroni correction: $\alpha^* = \alpha / m$
  - ✓ α : significance level.
  - ✓ m : number of hypothesis tests.

- Example: Bonferroni to test 3 hypotheses with p:
  - ✓ **H1: $p = 0.01$**
  - ✓ H2: $p = 0.02$
  - ✓ H3: $p = 0.03$
    - $\alpha^* = \alpha / m = 0.05 / 3 = 0.0167$

  => We'd need $p \leq 0.0167$ to declare significance.

Bonferroni Correction

Table 1b | Efficacy of surgical face masks in reducing respiratory virus frequency of detection and viral shedding in respiratory droplets and aerosols of symptomatic individuals with coronavirus, influenza virus or rhinovirus infection

	Droplet particles >5 µm			Aerosol particles ≤5 µm		
Virus type	**Without surgical face mask**	**With surgical face mask**	***P***	**Without surgical face mask**	**With surgical face mask**	***P***
	Detection of virus					
	No. positive/no. total (%)	**No. positive/no. total (%)**		**No. positive/no. total (%)**	**No. positive/no. total (%)**	
Coronavirus	3 of 10 (30)	0 of 11 (0)	0.09	**4 of 10 (40)**	**0 of 11 (0)**	**0.04**
Influenza virus	**6 of 23 (26)**	**1 of 27 (4)**	**0.04**	8 of 23 (35)	6 of 27 (22)	0.36
Rhinovirus	9 of 32 (28)	6 of 27 (22)	0.77	19 of 34 (56)	12 of 32 (38)	0.15
	Viral load ($\log_{10}$ virus copies per sample)					
	Median (IQR)	**Median (IQR)**		**Median (IQR)**	**Median (IQR)**	
Coronavirus	0.3 (0.3, 1.2)	0.3 (0.3, 0.3)	0.07	**0.3 (0.3, 3.3)**	**0.3 (0.3, 0.3)**	**0.02**
Influenza virus	**0.3 (0.3, 1.1)**	**0.3 (0.3, 0.3)**	**0.01**	0.3 (0.3, 3.0)	0.3 (0.3, 0.3)	0.26
Rhinovirus	0.3 (0.3, 1.3)	0.3 (0.3, 0.3)	0.44	1.8 (0.3, 2.8)	0.3 (0.3, 2.4)	0.12

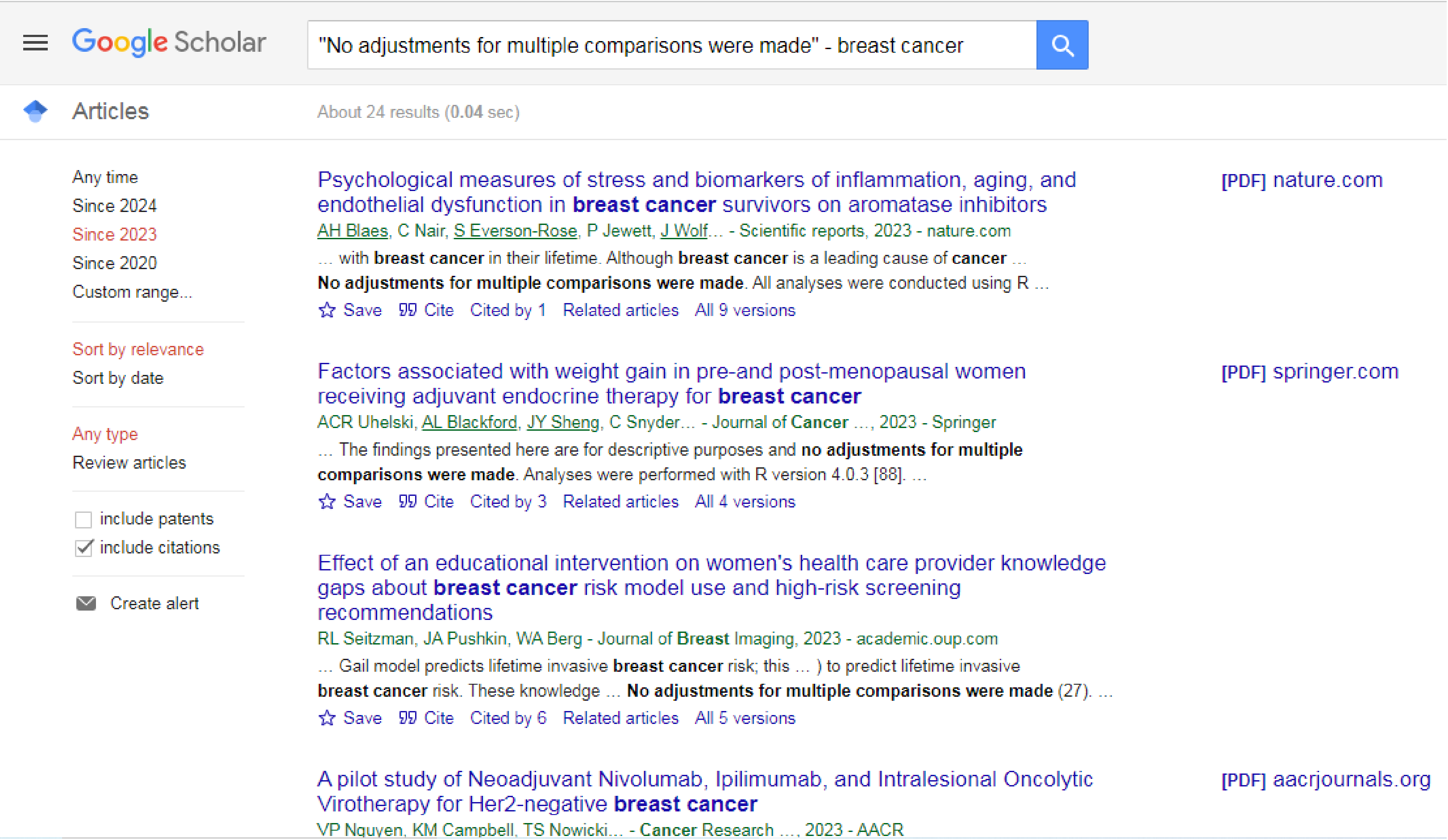
P values for comparing the frequency of respiratory virus detection between the mask intervention were obtained by two-sided Fisher's exact test and (two-sided) *P* values for mask intervention as predictor of $\log_{10}$ virus copies per sample were obtained by an unadjusted univariate Tobit regression model, which allowed for censoring at the lower limit of detection of the RT-PCR assay, with significant differences in bold. Undetectable values were imputed as 0.3 $\log_{10}$ virus copies per sample. IQR, interquartile range.

$\alpha^* = \alpha / m = 0.05 / 12 = 0.004$

=> We'd need p ≤ 0.004 to declare significance.

In article

- “No adjustments for multiple comparisons were made”.



In article

frontiers | Frontiers in Oncology

TYPE Original Research
PUBLISHED 06 September 2022
DOI 10.3389/fonc.2022.988794

Check for updates

OPEN ACCESS

EDITED BY
Vasiliki Gkretsi,
European University Cyprus, Cyprus

REVIEWED BY
Cecilia Battistelli,
Sapienza University of Rome, Italy
Yen-Nien Liu,
Taipei Medical University, Taiwan

*CORRESPONDENCE
Guan-Ming Kuang
kuanggm@hku-szh.org

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to

microRNA-145-5p inhibits prostate cancer bone metastatic by modulating the epithelial-mesenchymal transition

Bingfeng Luo[1†], Yuan Yuan[1†], Yifei Zhu[1], Songwu Liang[1], Runan Dong[1], Jian Hou[1], Ping Li[2], Yaping Xing[1], Zhenquan Lu[1], Richard Lo[1] and Guan-Ming Kuang[3*]

[1]Division of Urology, Department of Surgery, The University of Hong Kong-Shenzhen Hospital, Shenzhen, China, [2]Department of Pathology, The University of Hong Kong-Shenzhen Hospital, Shenzhen, China, [3]Department of Orthopedics and Traumatology, The University of Hong Kong-Shenzhen Hospital, Shenzhen, China

incubation for 1 h. For the migration/wound healing assays, 3 × 10^5 cells/well were grown in a 24-well plate, incubated for 16-18 h and cell monolayers scraped with a pipette tip to create a wound which was washed with PBS. After incubation for 24 h in culture medium, an inverted microscope with a digital camera was used to photograph wound closure. Colony formation was measured by resuspending the cells with 1ml medium and seeding a six-well plate with 500 cells per well. After 2 weeks, 6-well plates were fixed with paraformaldehyde (4%) for 30min at room temperature before washing with PBS, the addition of crystal violet staining and photographs taken under the microscope. Transwell assay was conducted to assess invasion.

TBS, proteins were visualized with an electroluminescence kit (ASPEN, Wuhan, China). The internal control was GAPDH.

Statistical analysis

Means ± SD of three independent experiments were presented, and statistical analysis was conducted using GraphPad v4.1 (CA, USA). Data were compared between groups using a two-tailed unpaired Student's t-test. A p-value of <0.05 was deemed statistically significant.

In article

Statistical analysis

Statistical analyses were performed using Pearson's Chi-squared test or Fisher's exact test to determine significant clinicopathological differences between EGFR expression in positive and negative tumors, between EGFR FISH-positive and FISH-negative tumors, and between tumors with and without EGFR mutations. These tests were also used to determine the association between EGFR protein expression, EGFR FISH results, and EGFR mutations. Bonferroni correction was performed to adjust for multiple comparisons, differences with $P < 0.05$/comparison times were considered significant.

EGFR mutations in lung adenocarcinomas

Eighty-five (63.9%) of the 133 cases showed EGFR mutations, which included 2 exon 18 G719X mutations (one also had an exon 20 S768I mutation), 39 exon 19 deletions, 4 exon 20 insertion mutations, 3 exon 20 S768I mutations (one also had an exon 18 G719X mutation), 35 exon 21 L858R mutations (one also had an exon 20 T790 M mutations), and 3 exon 21 L861Q mutation. After Bonferroni correction for 5 comparisons, $P < 0.01$ were considered significant, EGFR mutations were significantly associated with smoking status (non-smoking vs. smoking, $p = 0.008$), and were not associated with age, gender, lymph node metastasis or tumor stage ($p \geq 0.01$) (Table 1).

Liang Z, Zhang J, Zeng X, Gao J, Wu S, Liu T. Relationship between EGFR expression, copy number and mutation in lung adenocarcinomas. BMC cancer. 2010 Dec;10:1-9.

2. Selective reporting independent variable

Background

- **p-hacking**:
  - ✓ Multiple experimental groups vs 1 control group.
    - Example: Different Drug vs Control

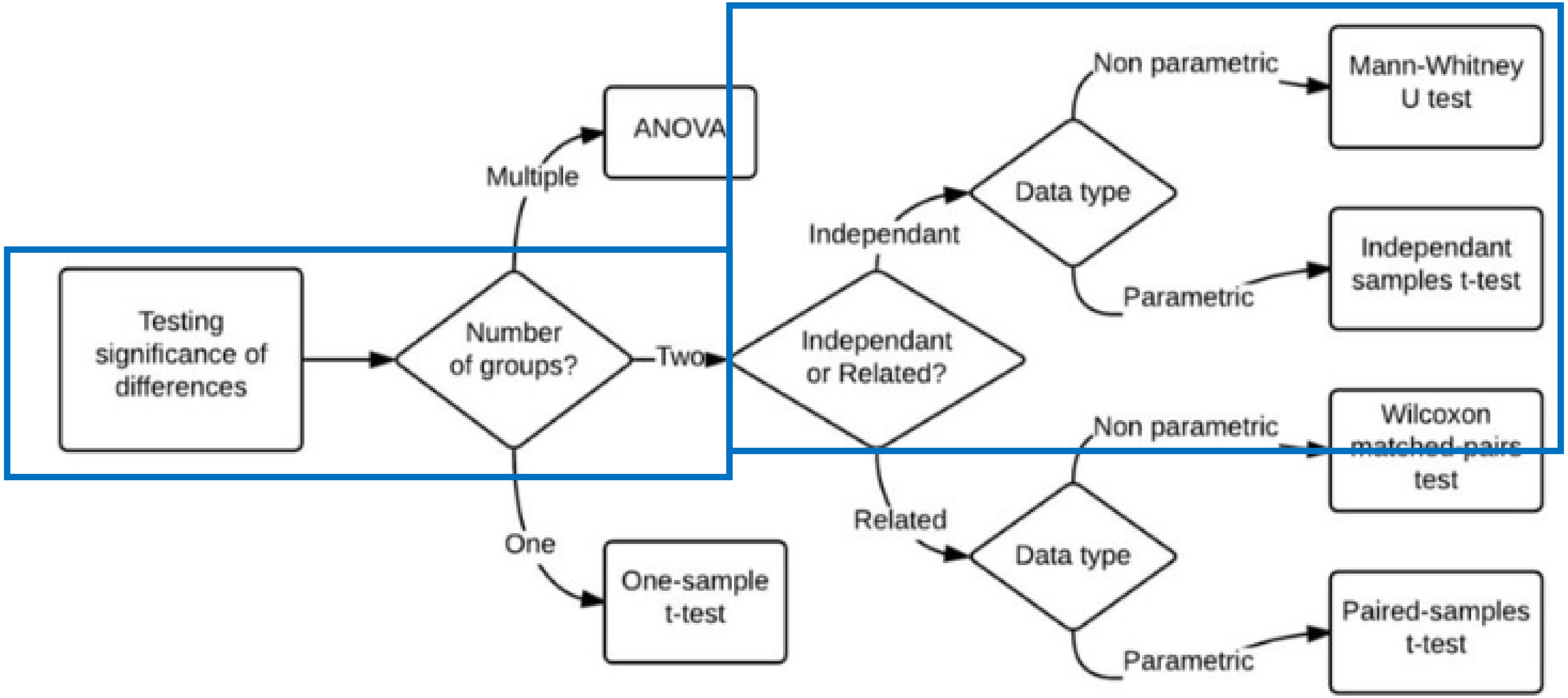
      Different Drug Concentrations vs Control
  - ✓ Compares all experimental groups to the control group.
  - ✓ Only report the significant results.

Background

Control	Drug 1	Drug 2
1147	1169	1009
1273	1323	1260
1216	1276	1143
1046	1240	1099
1108	1432	1385
1265	1562	1164

Lew M. Good statistical practice in pharmacology Problem 2. British journal of pharmacology. 2007 Oct;152(3):299-303.

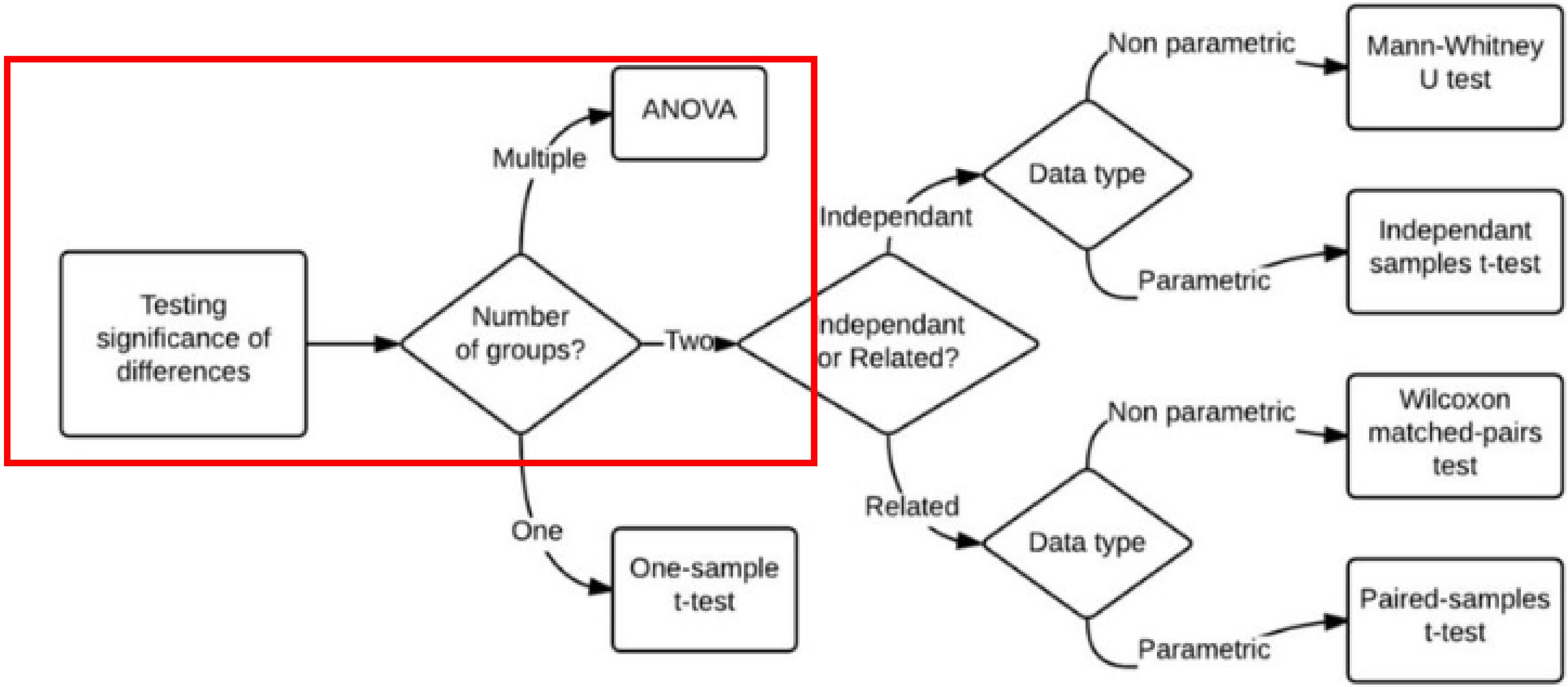
Decision tree for statistical analysis – 2 groups



Statistical analysis decision tree for testing significance of differences

Borghini YC. *An Assessment and Learning Analytics Engine for Games-based Learning* (Doctoral dissertation, University of the West of Scotland).

Decision tree for statistical analysis - 3 groups



Statistical analysis decision tree for testing significance of differences

Borghini YC. *An Assessment and Learning Analytics Engine for Games-based Learning* (Doctoral dissertation, University of the West of Scotland).

t-test vs ANOVA

		Independent t-test	ANOVA
Null Hypothesis		No difference between population's means. Ho: μ1= μ2	No difference between population's means. Ho: μ1= … = μk
Alternative hypothesis		Difference between 2 populations' means. H1: μ1 ≠ μ2	At least 2 group means are different from each other. H1: μ1 ≠ μ2 **or** μ1 ≠ μ3 **or** μ2 ≠ μ3 ….
Conclusion	$p > 0.05$	We **don't have enough evidence** to conclude that the difference is statistically significant.	We **don't have enough evidence** to conclude that the difference is statistically significant.
	$p \leq 0.05$	We **have enough evidence** to conclude that the difference is statistically significant.	There is a **significant effect** of independent variable **on levels of / according to** response variable.

t-test vs ANOVA

- After perform hypothesis test:
  - ✓ **Independent t-test** ⇨ Conclusion 2 groups
  - ✓ **ANOVA** ⇨ Which groups differ ??
- **Post Hoc Tests for ANOVA**
  - ✓ 1 vs 2
  - ✓ 1 vs 3
  - ✓ …
  - ✓ m vs n

Post Hoc Tests for ANOVA - R

- Multiple testing issue: P (At least 1 error in m tests) $= 1 - (1 - \alpha)^m$
- 2 approaches:
  - ✓ **Compare $p \leq \alpha^*$** — $\alpha^* = \alpha / m = 0.05 / 3 = 0.017.$
  - ✓ **Compare $p^* \leq \alpha$** — $p^* = p * m = p * 3$

Bonferroni

```
Pairwise comparisons using t
tests with pooled SD

data:  viagraData$libido and
viagraData$dose

          Placebo Low Dose
Low Dose  0.845   -
High Dose 0.025   0.196

P value adjustment method:
bonferroni
```

BH

```
Pairwise comparisons using t
tests with pooled SD

data:  viagraData$libido and
viagraData$dose

          Placebo Low Dose
Low Dose  0.282   -
High Dose 0.025   0.098

P value adjustment method: BH
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = libido ~ dose, data = viagraData)

$dose
                    diff        lwr      upr     p adj
Low Dose-Placebo     1.0 -1.3662412 3.366241 0.5162761
High Dose-Placebo    2.8  0.4337588 5.166241 0.0209244
High Dose-Low Dose   1.8 -0.5662412 4.166241 0.1474576
```

Output 10.11

Background

ANOVA

Control	Drug 1	Drug 2
1147	1169	1100
1273	1323	1125
	…	
1108	1276	1110
1265	1240	1000

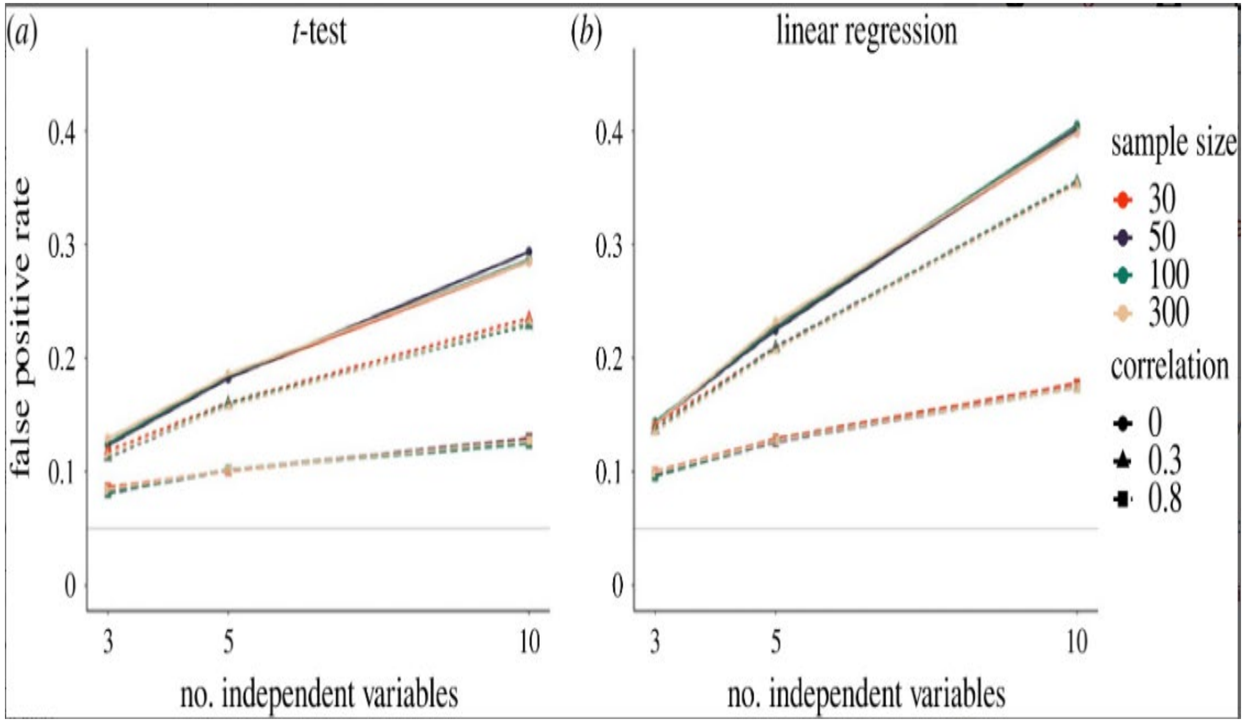
t-test →

Control	Drug 1
1147	1169
…	
1265	1240

Control	Drug 2
1147	1100
…	
1265	1000

2. Selective reporting the independent variable

- Assume using t-test.
- Sample size: not a protective factor.
- **FPR decrease** with:
  - ✓ **Less independent variables**.
  - ✓ **Higher correlation variables**.
- Severe effects in regression >> t-tests.



Number of independent variables indicates how many hypothesis tests were conducted (at maximum) to obtain a significant result.
The solid grey line: nominal α-level of 5%.
(a) FPR for the *t*-test.
(b) FPR for a univariate regression.

Conclusion

- Selective reporting DV
  - ✓ What is correlation
  - ✓ Multiple testing hypothesis issue

- Selective reporting IV
  - ✓ Post Hoc Tests for ANOVA

References

1. Stefan AM, Schönbrodt FD. Big little lies: A compendium and simulation of p-hacking strategies. Royal Society Open Science. 2023 Feb 8;10(2):220346.

Thank you for listening