Type I and Type II errors

Issue Multiple Hypothesis Testing

Presenter: Bui Minh Tri

Objective + Outline

- Objective

Type I and Type II errors - Issue Multiple Hypothesis Testing

- Outline

1. **4 situations - Statistical decision-making**
2. **Significance level (α)**
3. **β vs Power (1-β)**
4. **Relationship α and β**
5. **Issue with Multiple Testing of hypothesis**

4 situations

Statistical decision-making

Independent-means t-test

	Case N=503	Control N=493	p value
Age	45.0 [39.0;54.0]	51.0 [41.0;57.0]	<0.001
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

4 situations in statistical decision-making

In Independent-means t-test:

- Null Hypothesis: no difference between 2 populations' means.
  - ✓ Ho: $\mu_1 = \mu_2$

- Research Hypothesis: difference between 2 populations' means.
  - ✓ H1: $\mu_1 \neq \mu_2$

4 situations in statistical decision-making

- Compare p with 0.05:

  - ✓ **p ≤ .05 ➢ Reject Null Hypothesis** ➢ We have **enough evidence** to conclude that the difference between groups is statistically significant.

  - ✓ **p > .05 ➢ Failed to reject Null Hypothesis** ➢ We **don't have enough evidence** to conclude that the difference between groups is statistically significant.

4 situations in statistical decision-making

Decision of Test	REALITY Null Hypothesis H0 μ1 = μ2	
	TRUE	FALSE
Reject H0		
Not reject H0		

4 situations in statistical decision-making

Decision of Test	REALITY Null Hypothesis H0 μ1 = μ2	
	TRUE	FALSE
Reject H0	**Type I Error** False Positive Probability: α	
Not reject H0		

4 situations in statistical decision-making

Decision of Test	REALITY Null Hypothesis H0 $\mu 1 = \mu 2$	
	TRUE	FALSE
Reject H0	**Type I Error** False Positive Probability: α	
Not reject H0	Correct decision True Negative Probability: $1 - \alpha$	

4 situations in statistical decision-making

Decision of Test	REALITY Null Hypothesis H0 $\mu 1 = \mu 2$	
	TRUE	FALSE
Reject H0	**Type I Error** False Positive Probability: α	
Not reject H0	Correct decision True Negative Probability: $1 - \alpha$	**Type II Error** False Negative Probability: β

4 situations in statistical decision-making

Decision of Test	REALITY Null Hypothesis H0 $\mu 1 = \mu 2$	
	TRUE	FALSE
Reject H0	**Type I Error** False Positive Probability: α	Correct decision True Positive Probability: $1 - \beta$
Not reject H0	Correct decision True Negative Probability: $1 - \alpha$	**Type II Error** False Negative Probability: β

Significance level (α)

Significance level (α)

- Ho: $\mu_1 = \mu_2$
- H1: $\mu_1 \neq \mu_2$

✓ **p ≤ .05 ➢ Reject Null Hypothesis** ➢ We have **enough evidence** to conclude that the difference between groups is statistically significant.

✓ **p > .05 ➢ Failed to reject Null Hypothesis** ➢ We **don't have enough evidence** to conclude that the difference between groups is statistically significant.

Significance level (α)

- The probability that the observed difference could have occurred by chance.
  - ✓ **α = 5%:**
    - 5% probability that observed difference occurred by chance.

Significance level (α)

- The probability that the observed difference could have occurred by chance.
  - ✓ **α = 5%:**
    - 5% probability that observed difference occurred by chance.
    - 5% risk of concluding that a difference exists when there is none.

Significance level (α)

- The probability that the observed difference could have occurred by chance.
  - ✓ **α = 5%:**
    - 5% probability that observed difference occurred by chance.
    - 5% risk of concluding that a difference exists when there is none.
    - 5% risk of **false positive**.

Significance level (α)

- The probability that the observed difference could have occurred by chance.
  - ✓ **α = 5%:**
    - 5% probability that observed difference occurred by chance.
    - 5% risk of concluding that a difference exists when there is none.
    - 5% risk of **false positive**.

- Predetermined threshold to make a decision about the null hypothesis.
  - ✓ **A cut off point**: **$p \leq \alpha$**
    - ➢ Reject Null Hypothesis (Ho: $\mu_1 = \mu_2$).
    - ➢ Accept α risk of false positive.

Note 1: p vs α

	p	α
Definition	The probability of observing these data or more extreme data, if the null hypothesis is true.	Predetermined threshold that you use to make a decision about the null hypothesis (rejecting null hypothesis).
Interpretation	**Strength of evidence** against the null hypothesis.	**$p \leq \alpha$ ⇨ Reject Null Hypothesis** **$p > \alpha$ ⇨ Fail to reject Null Hypothesis**
When?	Result of statistical test.	Predetermined value.

Note 2: Strength of evidence

- **α value: 0.05**

- **No sharp distinction between "significant" and "not significant"** results, only increasing the **strength of evidence** against null hypothesis
  - ✓ **0.049 ≤ 0.05** vs **0.051 > 0.05**.

Note 2: Strength of evidence

- **α value: 0.05**

- **No sharp distinction between "significant" and "not significant"** results, only increasing the **strength of evidence** against null hypothesis
  - ✓ **0.049 ≤ 0.05** vs **0.051 > 0.05**.
  - ✓ Observed data not provide **strong enough evidence** to reject the null hypothesis.
  - ✓ Could still be a real effect or difference, but it might be smaller than the study was able to detect.

Note 2: Strength of evidence

- p-values: continuum and provide a relative measure of strength of evidence:
  - ✓ $p \geq 0.1$ insufficient evidence
  - ✓ $p < 0.1$ weak evidence
  - ✓ **$p < 0.05$ moderate evidence**
  - ✓ $p < 0.01$ strong evidence
  - ✓ $p < 0.001$ very strong evidence

β

Power (1-β)

4 situations in statistical decision-making

Decision of Test	REALITY Null Hypothesis H0 $\mu 1 = \mu 2$	
	TRUE	FALSE
Reject H0	**Type I Error** False Positive Probability: α	Correct decision True Positive Probability: $1 - \beta$
Not reject H0	Correct decision True Negative Probability: $1 - \alpha$	**Type II Error** False Negative Probability: β

β vs 1-β

- **β**: Probability of making a type II error - failing to reject the null hypothesis when it is actually false.

- **Power: 1 – β:** Probability of observing an effect in the sample.

- **β = 0.2 (20%) ⇔ Power = 0.8 (80%)**
  - ✓ 20% False Negative.
  - ✓ If there are true effects to be found in 100 different studies with 80% power, only 80 out of 100 statistical tests will actually detect them.

Relationship α and β

Relationship α and β

Decision of Test	REALITY Null Hypothesis H0 μ1 = μ2	
	TRUE	FALSE
Reject H0	**Type I Error** False Positive Probability: α	Correct decision True Positive Probability: 1 – β
Not reject H0	Correct decision True Negative Probability: 1 – α	**Type II Error** False Negative Probability: β

Relationship α and β

- Ideally to eliminate false-positive and false-negative results ?
  - ✓ **α = 0 ⇔ False positive = 0**
  - ✓ **β = 0 ⇔ False negative = 0**

Relationship α and β

G*Power: a tool to

- Compute statistical power analyses.
- Compute effect sizes.
- Display graphically the results of power analyses.
- https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower

Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses

F Faul, E Erdfelder, A Buchner, AG Lang

Behavior research methods, 2009 · Springer

Abstract

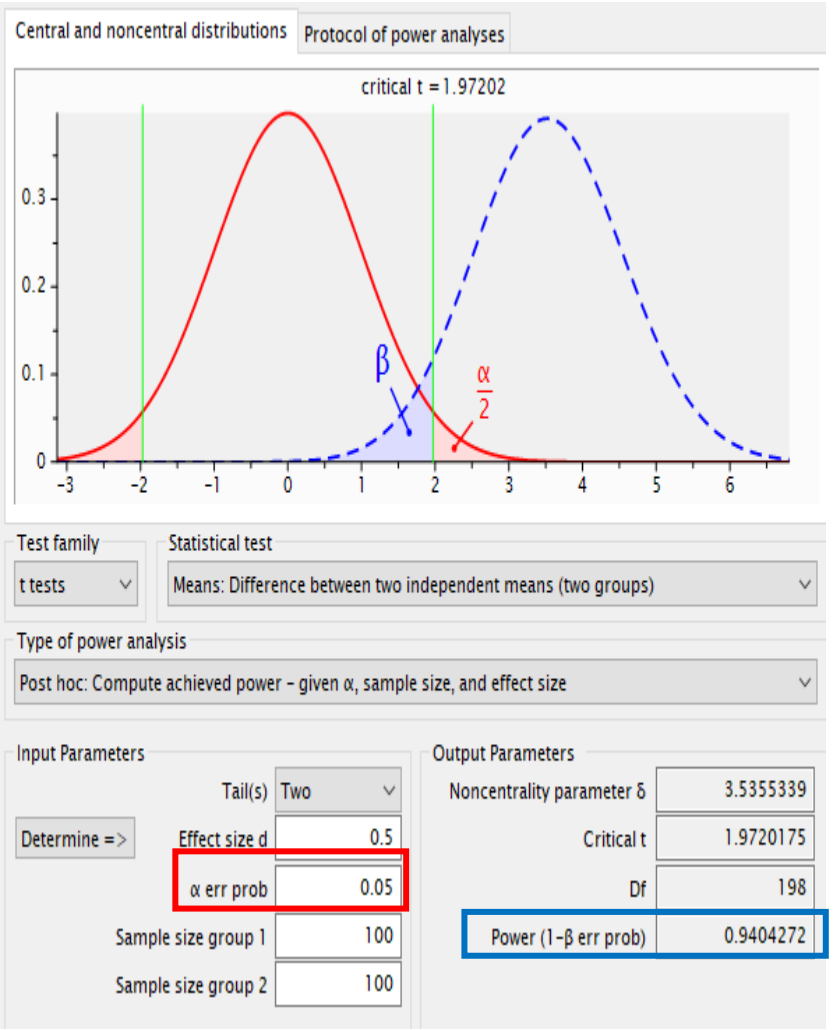
G*Power is a free power analysis program for a variety of statistical tests. We present extensions and improvements of the version introduced by Faul, Erdfelder, Lang, and Buchner (2007) in the domain of correlation and regression analyses. In the new version, we have added procedures to analyze the power of tests based on (1) single-sample tetrachoric correlations, (2) comparisons of dependent correlations, (3) bivariate linear regression, (4) multiple linear regression based on the random predictor model, (5) logistic regression, and (6) Poisson regression. We describe these new features and provide a brief introduction to their scope and handling.

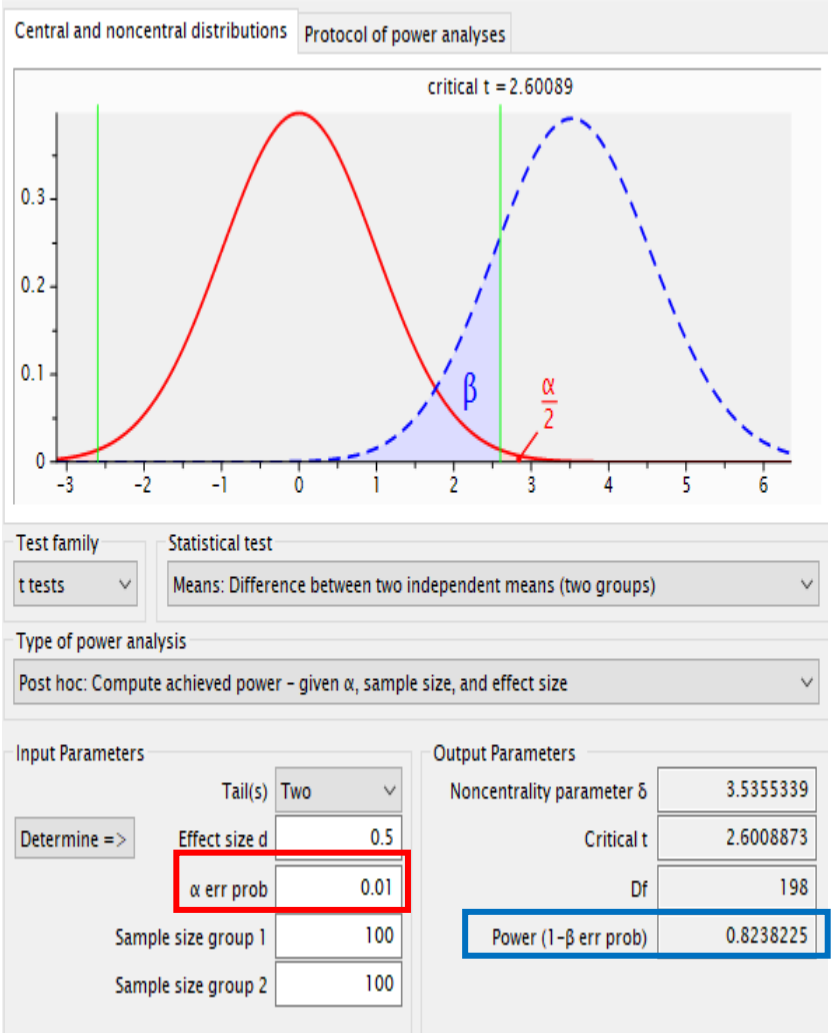
Springer

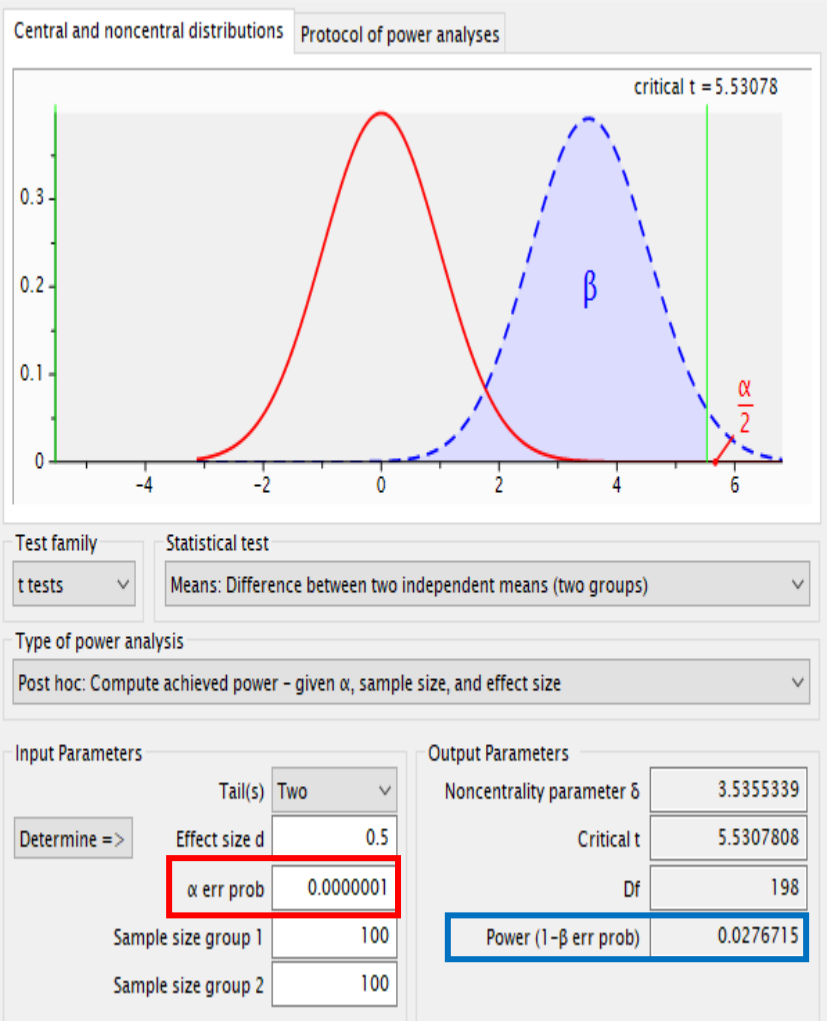
SHOW LESS ^

Save Cite Cited by 34200 Related articles All 23 versions

Relationship α and β







Relationship α and β

- Ideally to eliminate false-positive and false-negative results ?
  - ✓ **α = 0 ⇔ False positive = 0.**
  - ✓ **β = 0 ⇔ False negative = 0.**

- **α decrease ⇔ β increase**.

- Many studies:
  - ✓ **α = 0.01 or 0.05**
  - ✓ **β = 0.1 (Power = 0.9) or 0.20 (Power = 0.80)**

Issue of Multiple testing of hypothesis

Multiple testing issue

- **α = 5%:** Risk of false positive rate **for 1 test** = 5%.
- Multiple hypothesis tests inflated the risk of type I error – **Family-wise / Experiment-wise error rate (FWER).**

Multiple testing issue

BRIEF COMMUNICATION
https://doi.org/10.1038/s41591-020-0843-2
nature medicine
Check for updates

Respiratory virus shedding in exhaled breath and efficacy of face masks

Nancy H. L. Leung[1], Daniel K. W. Chu[1], Eunice Y. C. Shiu[1], Kwok-Hung Chan[2], James J. McDevitt[3], Benien J. P. Hau[1,4], Hui-Ling Yen[1], Yuguo Li[5], Dennis K. M. Ip[1], J. S. Malik Peiris[1], Wing-Hong Seto[1,6], Gabriel M. Leung[1], Donald K. Milton[7,8] and Benjamin J. Cowling[1,8]

We identified seasonal human coronaviruses, influenza viruses and rhinoviruses in exhaled breath and coughs of children and adults with acute respiratory illness. Surgical face masks significantly reduced detection of influenza virus RNA in respiratory droplets and coronavirus RNA in aerosols, with a trend toward reduced detection of coronavirus RNA in respiratory droplets. Our results indicate that surgical face masks could prevent transmission of human coronaviruses and influenza viruses from symptomatic individuals.

Respiratory virus infections cause a broad and overlapping spectrum of symptoms collectively referred to as acute respiratory virus illnesses (ARIs) or more commonly the 'common cold'. Although mostly mild, these ARIs can sometimes cause severe disease and medically attended ARIs and determining the potential efficacy of surgical face masks to prevent respiratory virus transmission.

Results

We screened 3,363 individuals in two study phases, ultimately enrolling 246 individuals who provided exhaled breath samples (Extended Data Fig. 1). Among these 246 participants, 122 (50%) participants were randomized to not wearing a face mask during the first exhaled breath collection and 124 (50%) participants were randomized to wearing a face mask. Overall, 49 (20%) voluntarily provided a second exhaled breath collection of the alternate type.

Infections by at least one respiratory virus were confirmed by reverse transcription PCR (RT–PCR) in 123 of 246 (50%) partici-

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)	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.

Family-wise / Experiment-wise error rate (FWER)

- If perform **m hypothesis independent tests**, **the probability at least 1 false positive** ?
  - ✓ P (Making Type I error) $= \alpha$
  - ✓ 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%

FWER

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

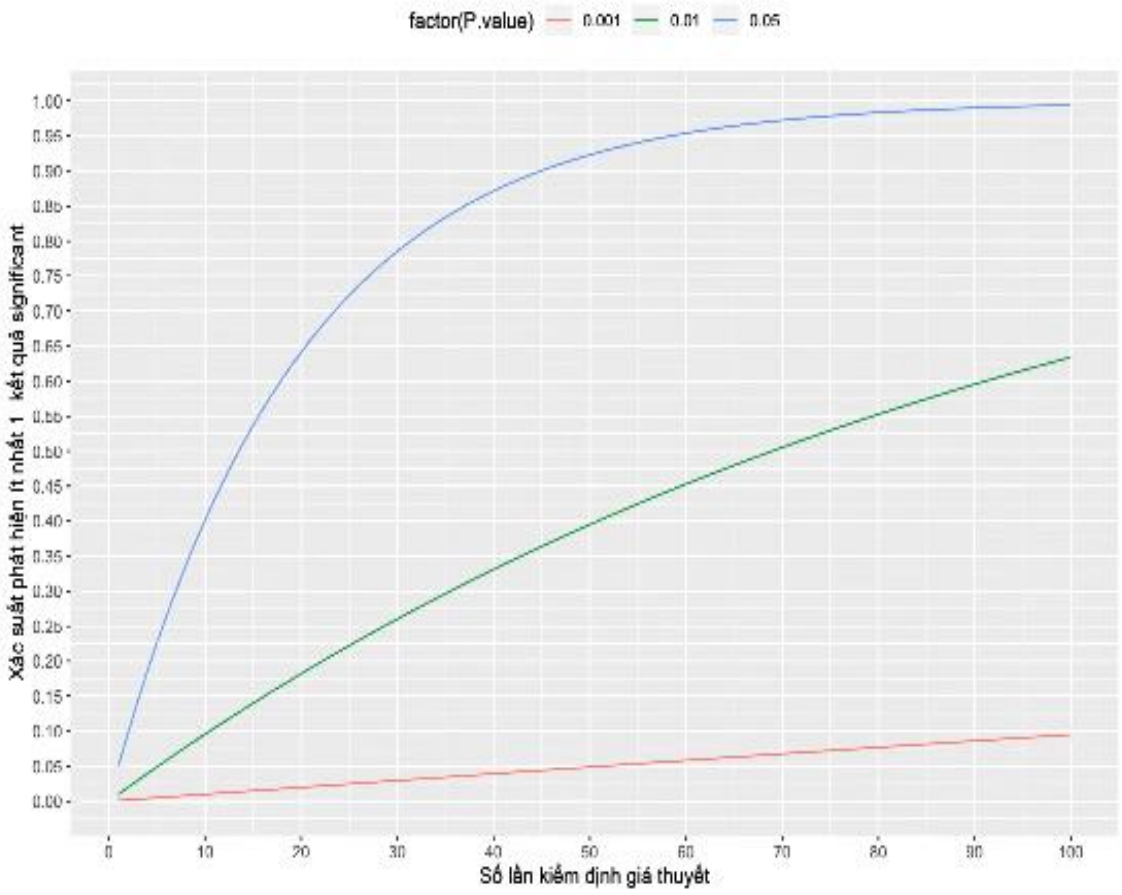
	Droplet particles >5 µm			Aerosol particles ≤5 µm		
Virus type	**Without surgical face mask**	**With surgical face mask**	***P***	**Without surgical face mask**	**With surgical face mask**	***P***
	Detection of virus					
	No. positive/no. total (%)	**No. positive/no. total (%)**		**No. positive/no. 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.

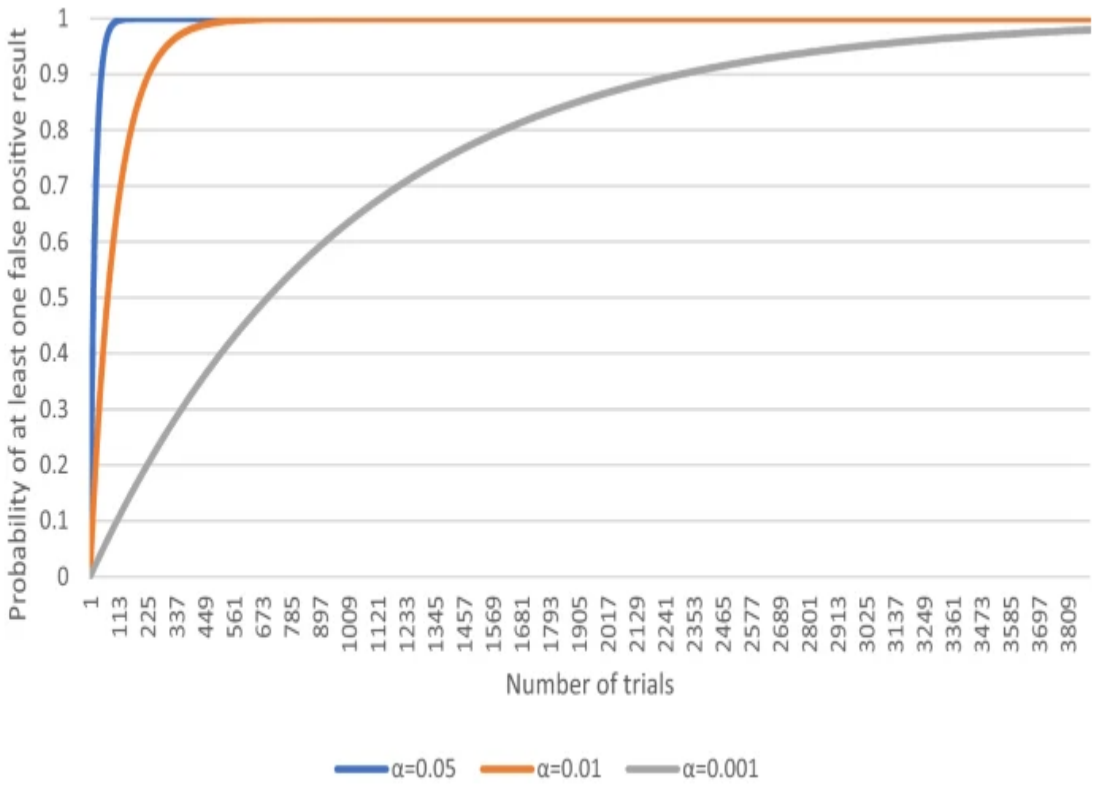
P (Making at least 1 error in 12 tests):

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

FWER



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.

FWER – Correction

- **Single Step**: equivalent adjustments made to each p-value.
- **Sequential**: adaptive adjustment made to each p-value.

Single Step – Bonferroni Concept

- Simple method to maintain overall Type I error rate (α) when performing m independent hypothesis tests.
- When ?
  - ✓ Multiple ‘t’ tests / Mann-Whitney
  - ✓ Post-hoc test after ANOVA / Kruskal-Wallis test
  - ✓ Pearson's ‘r’
  - ✓ Chi-square / Contingency table test

Single Step – Bonferroni Concept

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

Single Step – Bonferroni Concept

- 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.

Single Step – Bonferroni Example 1

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

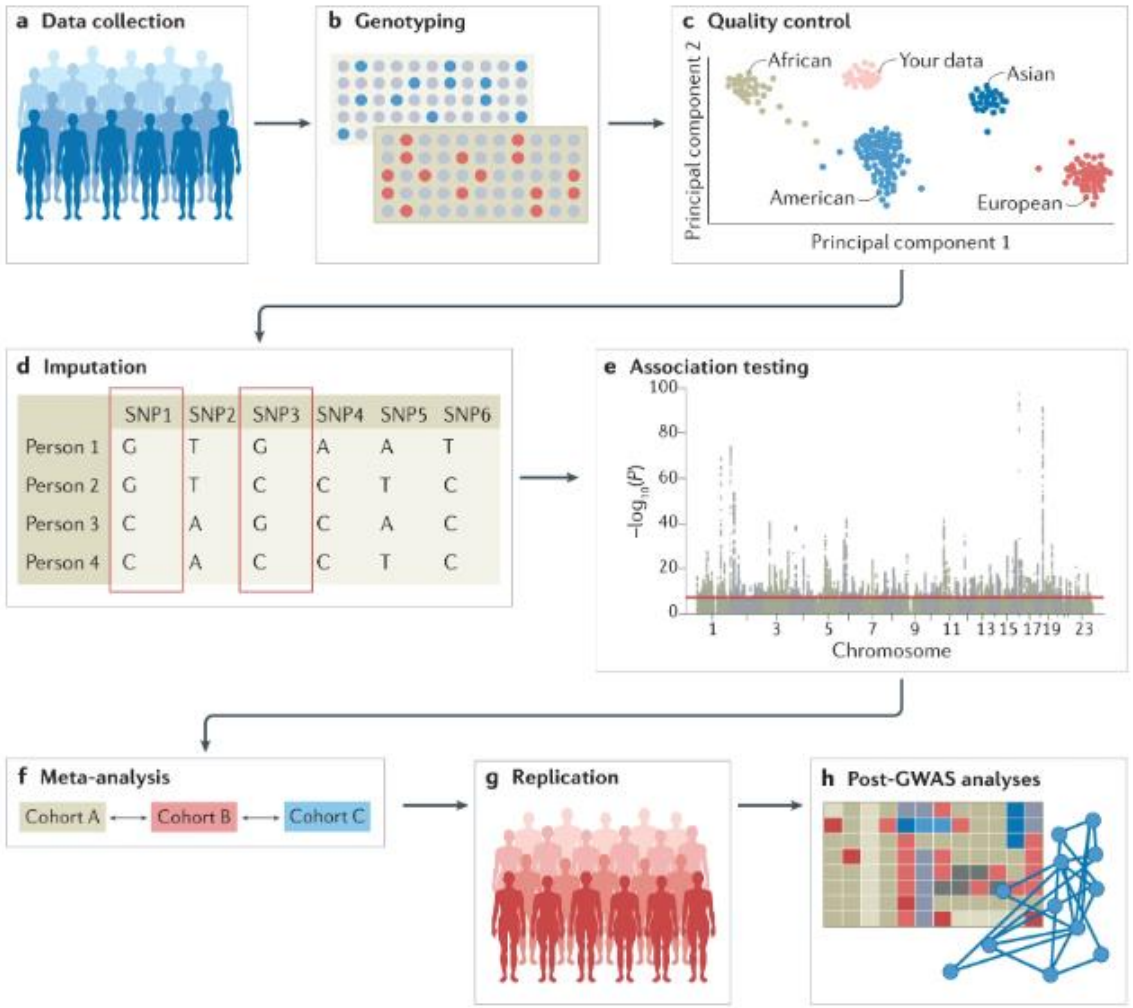
P values for comparing the frequency of respiratory virus detection between the mask intervention were obtained by two-sided Fisher's exact test and (two-sided) *P* values for mask intervention as predictor of $\log_{10}$ virus copies per sample were obtained by an unadjusted univariate Tobit regression 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 $p \leq 0.004$ to declare significance.

Single Step – Bonferroni Example 2

- Testing **millions of associations** between individual genetic variants and a phenotype of interest
- Multiple-testing threshold to avoid false positives:
  - ✓ Bonferroni testing threshold: $P < 0.05/ 10^6 = 5 \times 10^{-8}$



Overview of steps for conducting GWAS

Uffelmann E, Huang QQ, Munung NS, De Vries J, Okada Y, Martin AR, Martin HC, Lappalainen T, Posthuma D. Genome-wide association studies. Nature Reviews Methods Primers. 2021 Aug 26;1(1):59.

Single Step - Controversy over Bonferroni

- **Benefits:**
  - ✓ Controls FWER: ↓ Type I error risk (False Positive) .
  - ✓ Simple + easy to understand.

Single Step - Controversy over Bonferroni

- **Benefits:**
  - ✓ Controls FWER: ↓ Type I error risk (False Positive) .
  - ✓ Simple + easy to understand.
- **Drawbacks:**
  - ✓ ↓ Type I error (False Positive) ⇨ ↑ Type II error (False Negative).
  - ✓ Better for independence: all tests are independent of each other.
  - ✓ Number of tests performed ?
    - All tests in a report or a subset of them.
    - Tests performed but not included in the report.
    - Tests from the same data included in other reports.
  - ✓ **Treating all tests equally regardless of their importance or relevance**.

Sequential - Holm-Bonferroni

- Holm-Bonferroni correction: $\alpha^* = \alpha / (m - i + 1)$
  - ✓ α : significance level.
  - ✓ m : number of hypothesis tests.
  - ✓ i : rank number of pair (by degree of significance).

- Example: Holm-Bonferroni to test 3 hypotheses with p:
  - ✓ H1: p = 0.01
  - ✓ H2: p = 0.02
  - ✓ H3: p = 0.03

Sequential - Holm-Bonferroni

- **Step 1: Order p from smallest to greatest:**
  - ✓ H1: p = 0.01
  - ✓ H2: p = 0.02
  - ✓ H3: p = 0.03
- **Step 2: α1* for 1st rank. Compare p-value to α1***:
  - ✓ H1: p = 0.01 < α1* = 0.05 / (3–1+1) = 0.0167 => **Reject Null Hypothesis**.
- **Step 3: α2* for 2nd rank:**
  - ✓ H2: p = 0.02 < α2* = 0.05 / (3–2+1) = 0.025 => **Reject Null Hypothesis.**
- **Step 4: α3* for 3rd rank:**
  - ✓ H2: p = 0.03 < α3* = 0.05 / (3–3+1) = 0.05 => **Reject Null Hypothesis.**
- **Note: The test stops when you reach the first non-rejected hypothesis. All subsequent hypotheses are non-significant.**

Bonferroni or Holm-Bonferroni

Bonferroni	Holm-Bonferroni
Treat significance level **equally** for all tests.	Adjusts significance level - **order of p-values**.
High conservative. (less likely to reject null hypothesis)	**Less conservative**.
Not suitable for large number of tests.	Suitable for large number of tests.
$\alpha^* = \alpha / m$	$\alpha^* = \alpha / (m - i + 1)$
Example: **H1: p = 0.01** H2: p = 0.02 H3: p = 0.03 $\alpha^* = \alpha / m = 0.05 / 3 = 0.0167$	Example: **H1: p = 0.01** $\alpha 1^* = 0.05 / 3 = 0.0167$ **H2: p = 0.02** $\alpha 2^* = 0.05 / 2 = 0.025$ **H3: p = 0.03** $\alpha 3^* = 0.05 / 1 = 0.05$ Note: Stop when reach 1st **non-rejected hypothesis.**

False Discovery Rate (FDR)

- **FWER** control the probability of falsely rejecting **any** null hypothesis.
- But with **large number of test** ⇨ α* too low ⇨ **very low chance reject null hypothesis - super conservative**.
- Instead we can control **False Discovery Rate (FDR).**

False Discovery Rate (FDR)

Decision of Test	REALITY Null Hypothesis H0 μ1 = μ2	
	TRUE	**FALSE**
Reject H0	**Type I Error** False Positive Probability: α	Correct decision True Positive Probability: 1 – β
Not reject H0	Correct decision True Negative Probability: 1 – α	**Type II Error** False Negative Probability: β

FDR: The proportion of incorrect rejection of a hypothesis.

FDR = FP / (FP+TP) = Number of false rejection / Total number of rejection.

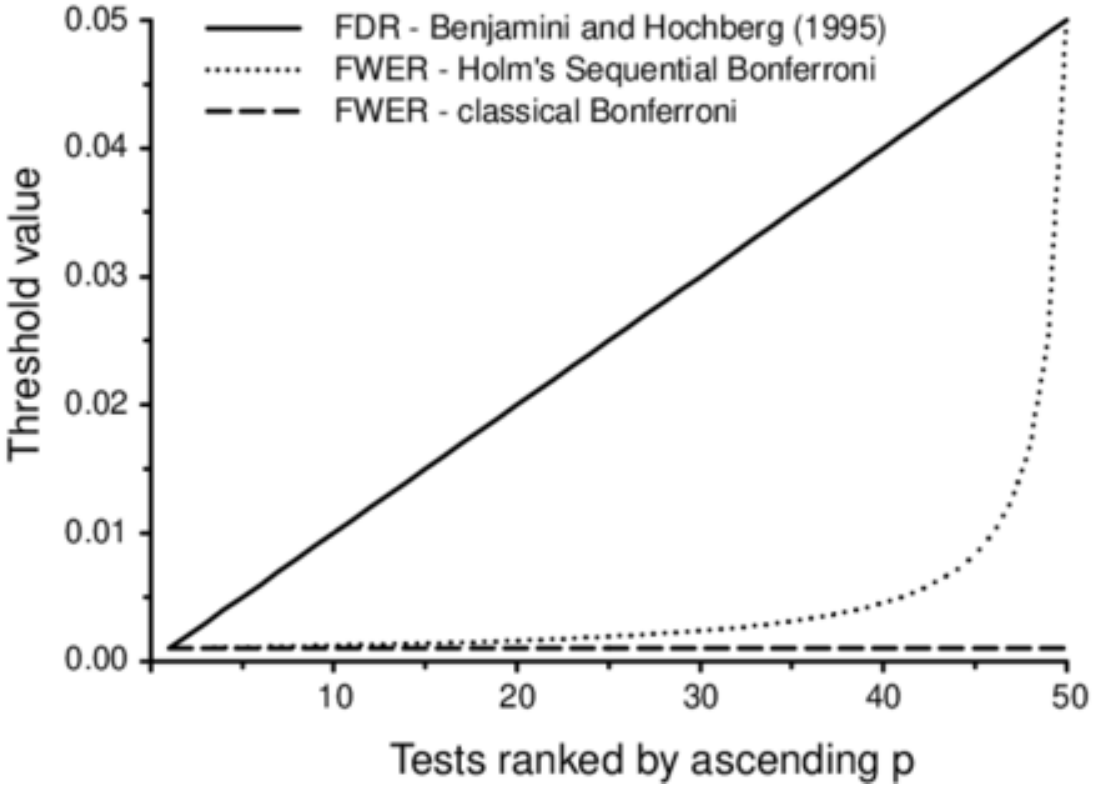
FWER vs FDR

If I conduct 1,000 hypothesis tests:

- FWER = 5%: Any individual test with a p-value ≤ 0.05 / 1,000 would be considered statistically significant.

- FDR = 5%: 100 tests are statistically significant ⇨ Expect up to 5 of significant results to be false positives.

FWER vs FDR



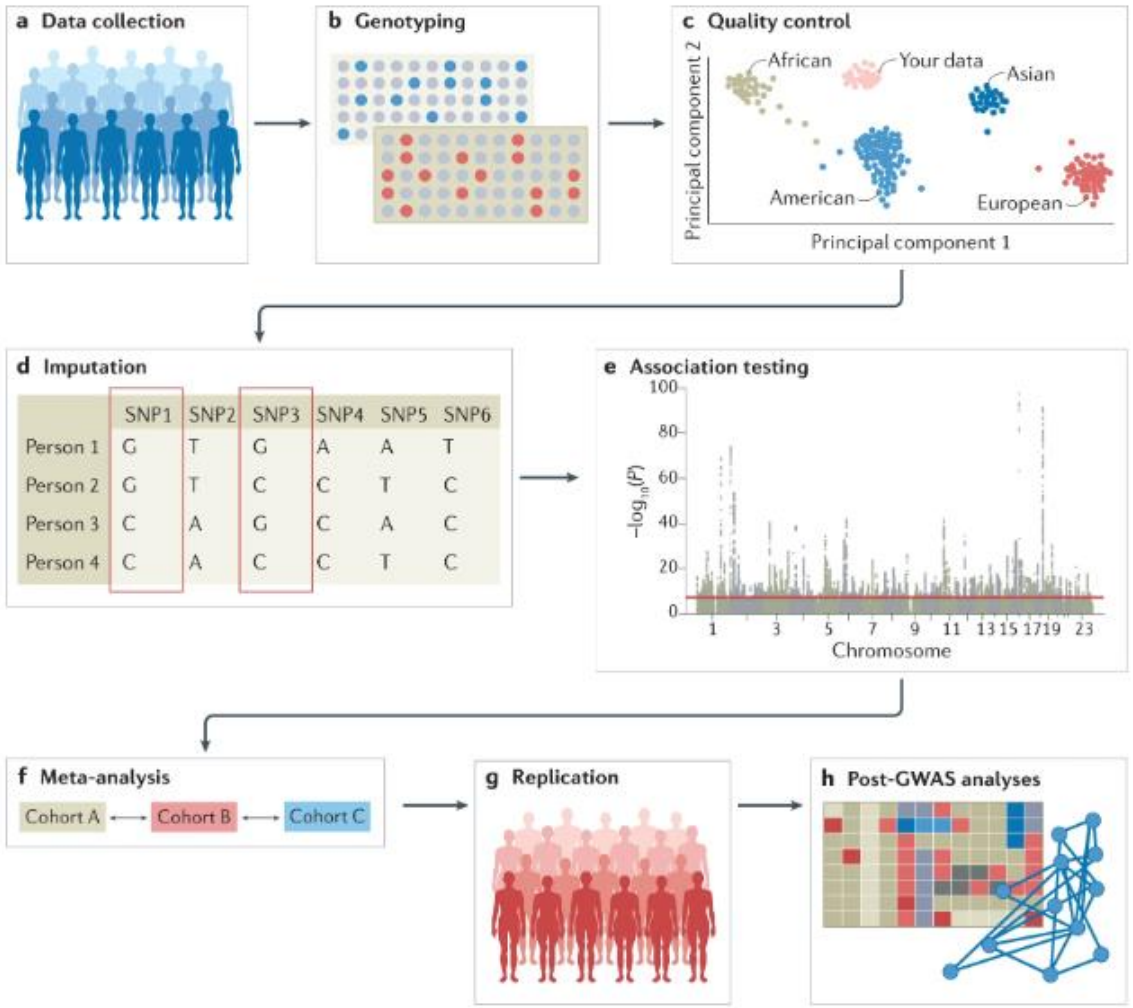
Tests ranked by p value	Bonferroni	Sequential Bonferroni	Benjamini and Hochberg
1	α / m	α / m	α / m
2	α / m	$\alpha / (m - 1)$	$2\alpha / m$
3	α / m	$\alpha / (m - 2)$	$3\alpha / m$
i	α / m	$\alpha / (m - i + 1)$	$i\,\alpha / m$
m	α / m	α	α

Comparison of threshold p values when 50 tests are performed

Verhoeven KJ, Simonsen KL, McIntyre LM. Implementing false discovery rate control: increasing your power. Oikos. 2005 Mar;108(3):643-7.

FWER vs FDR - Example GWAS

- Testing **millions of associations** between individual genetic variants and a phenotype of interest
- Multiple-testing threshold to avoid false positives:
  - ✓ Bonferroni testing threshold: $P < 0.05/ 10^6 = 5 \times 10^{-8}$
  - ✓ False discovery rate of $0.05/10^6$.



Overview of steps for conducting GWAS

Uffelmann E, Huang QQ, Munung NS, De Vries J, Okada Y, Martin AR, Martin HC, Lappalainen T, Posthuma D. Genome-wide association studies. Nature Reviews Methods Primers. 2021 Aug 26;1(1):59.

Benjamini-Hochberg

- $\alpha^* = FDR * (i/m)$
  - ✓ i: rank of p-value.
  - ✓ m: total number of tests.
  - ✓ FDR: your chosen false discovery rate.

- Example: FDR =0.05; H1: p = 0.01; H2: p = 0.03; H3: p = 0.04
  - ✓ H1: p = 0.01 < $\alpha 1^* = 0.05 * 1/3 = 0.0167$ => **Reject Null Hypothesis.**
  - ✓ H2: p = 0.03 < $\alpha 2^* = 0.05 * 2/3 = 0.033$ => **Reject Null Hypothesis.**
  - ✓ H3: p = 0.04 < $\alpha 3^* = 0.05 * 3/3 = 0.05$ => **Reject Null Hypothesis.**

- **Note: The test stops when you reach the first non-rejected hypothesis. All subsequent hypotheses are non-significant.**

FWER vs FDR

	FWER	FDR
Definition	Controls the **probability of making at least one false positive among all the tests or comparisons**. ⇨ Maintain **overall rate of false positives across all tests.**	Controls the **proportion of false positives among the rejected hypotheses.** ⇨ Focuses on **expected proportion of false positives among significant results.**
Formula	$\alpha^* = \alpha / m$	$\alpha^* = FDR * (i/m)$
Example	**H1: p = 0.01** $\alpha^* = 0.05 / 3 = 0.0167$ H2: p = 0.03 H3: p = 0.04	**H1: p = 0.01** $\alpha_1^* = 0.05 * 1/3 = 0.0167$ **H2: p = 0.03** $\alpha_2^* = 0.05 * 2/3 = 0.033$ **H3: p = 0.04** $\alpha_3^* = 0.05 * 3/3 = 0.05$
Trade-off	↓ false positives but ↑ false negatives. Can lead to many missed findings.	Balance between false positives - negatives. Can ↑ false positives cases.

Take-home messages

- p-values: continuum and provide a relative measure of strength of evidence:
  - ✓ $p \geq 0.1$ insufficient evidence
  - ✓ $p < 0.1$ weak evidence
  - ✓ **$p < 0.05$ moderate evidence**
  - ✓ $p < 0.01$ strong evidence
  - ✓ $p < 0.001$ very strong evidence
- **α (False positive) ↓ ⇔ β (False negative) ↑.**
- Multiple hypothesis tests inflated the risk of type I error:
  - ✓ FWER: Bonferroni / Holm-Bonferroni
  - ✓ FDR: Benjamini-Hochberg

References

1. Armstrong RA. When to use the B onferroni correction. Ophthalmic and Physiological Optics. 2014 Sep;34(5):502-8.
2. Diez DM, Barr CD, Cetinkaya-Rundel M. OpenIntro statistics. Boston, MA, USA:: OpenIntro; 2012.
3. Dubitzky W, Wolkenhauer O, Cho KH, Yokota H, editors. Encyclopedia of systems biology. New York, NY, USA:: Springer; 2013 Aug 17.
4. Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. Behavior Research Methods, 41, 1149-1160.
5. Ganesh S, Cave V. P-values, p-values everywhere!. New Zealand Veterinary Journal. 2018 Mar 4;66(2):55-6.
6. Kaur P, Stoltzfus J. Type I, II, and III statistical errors: A brief overview. International Journal of Academic Medicine. 2017 Jul 1;3(2):268-70.
7. 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.

References

8. Riffenburgh RH. Statistics in medicine. Academic press; 2012 Aug 13.

9. Serdar CC, Cihan M, Yücel D, Serdar MA. Sample size, power and effect size revisited: simplified and practical approaches in pre-clinical, clinical and laboratory studies. Biochemia medica. 2021 Feb 15;31(1):27-53.

10. Verhoeven KJ, Simonsen KL, McIntyre LM. Implementing false discovery rate control: increasing your power. Oikos. 2005 Mar;108(3):643-7.

11. https://nguyenvantuan830970966.files.wordpress.com/2020/10/21b6c-86994c_bdba554b17d74140b49d284f2a2b8091.pdf

12. https://gtpb.github.io/ABSTAT18/assets/Day_3/Slides_09_Multiple_Testing.pdf

13. https://www.youtube.com/watch?v=4RPUrwzgO6c