(5) Univariable Analysis of Categorical Data

Dr. Wan Nor Arifin

Biostatistics and Research Methodology Unit
Universiti Sains Malaysia
wnarifin@usm.my / wnarifin.github.io

Last update: Jul 16, 2023

Outlines

- Introduction
- Chi-squared test
- Fisher’s exact test
- McNemar’s test

Learning outcomes

- Understand the concept of non-parametric test
- Familiarize with selected non-parametric tests for categorical variables
- Understand and able to interpret the results of the selected non-parametric tests

Introduction

Non-parametric Test

- Statistical test that:
  - Distribution free, no assumptions about the distribution of the data e.g. normality, equality of variances
  - No specific population parameters to be tested, e.g. mean
  - Typically categorical; nominal or ordinal data
  - e.g. observed frequencies for categories in a sample number of smokers by gender etc

Non-parametric Test

- Statistical test that (cont.):
  - More flexible, can perform analysis when assumptions for parametric not fulfilled.
  - e.g. data not normally distributed.
  - LESS powerful than parametric test.

Non-parametric Test

- Non-parametric tests used for testing association for categorical outcomes:
  - Two categorical variables (two or more categories), one measurement: Chi-squared test, Fisher’s exact test
  - One categorical variable (two categories), two repeated measurements: McNemar’s test

Chi-squared Test

Chi-squared Test

- Purpose: Test the association between two categorical variables
- Procedure:
  - It compares the <u>observed</u> cell counts VS <u>expected</u> cell counts
  - If they differ substantially - association

Chi-squared Test

- Assumptions:
  - Only < 20% cells with expected count < 5
  - No expected counts < 1

Chi-squared Test

Research objective:

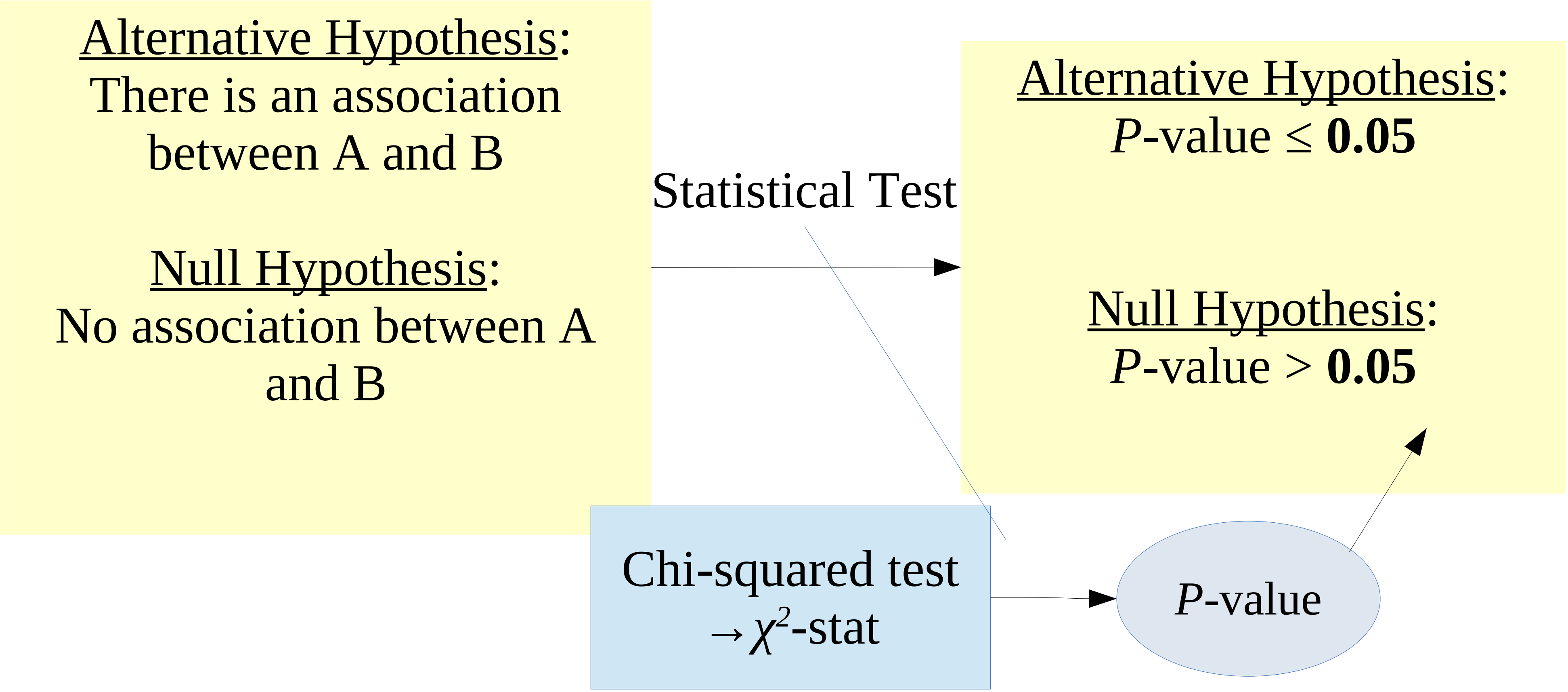
To test the association between A and B

Research question:

Is there an association between A and B?

Chi-squared Test

RQ: Is there an association between A and B?



<u>Alternative Hypothesis</u>:
There is an association between A and B

<u>Null Hypothesis</u>:
No association between A and B

Statistical Test

<u>Alternative Hypothesis</u>:
P-value ≤ **0.05**

<u>Null Hypothesis</u>:
P-value > **0.05**

Chi-squared test
→ χ^2-stat

P-value

Example

- Sample size: 200
- Variables:
  - Smoking: smoking / no smoking
  - Cancer: lung cancer / no lung cancer

Cross-tabulation

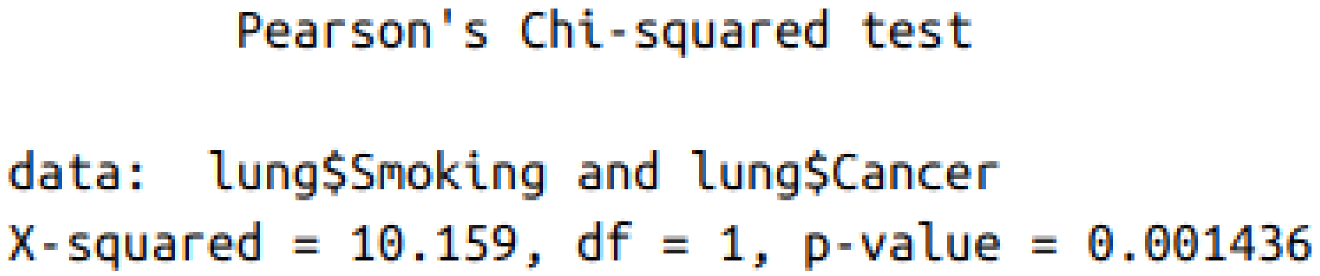
Smoking	Lung Cancer	
	Yes	No
Yes	20 (62.5%)	12 (37.5%)
No	55 (32.7%)	113 (67.3%)

Expected Count

Smoking	Lung Cancer		Sub-total
	Yes	No	
Yes	20 (32*75/200 = 12)	12 (32*125/200 = 20)	32
No	55 (168*75/200 = 63)	113 (168*125/200 = 105)	168
Sub-total	75	125	200

No expected count < 5

Results



```
        Pearson's Chi-squared test

data:  lung$Smoking and lung$Cancer
X-squared = 10.159, df = 1, p-value = 0.001436
```

P-value

Results

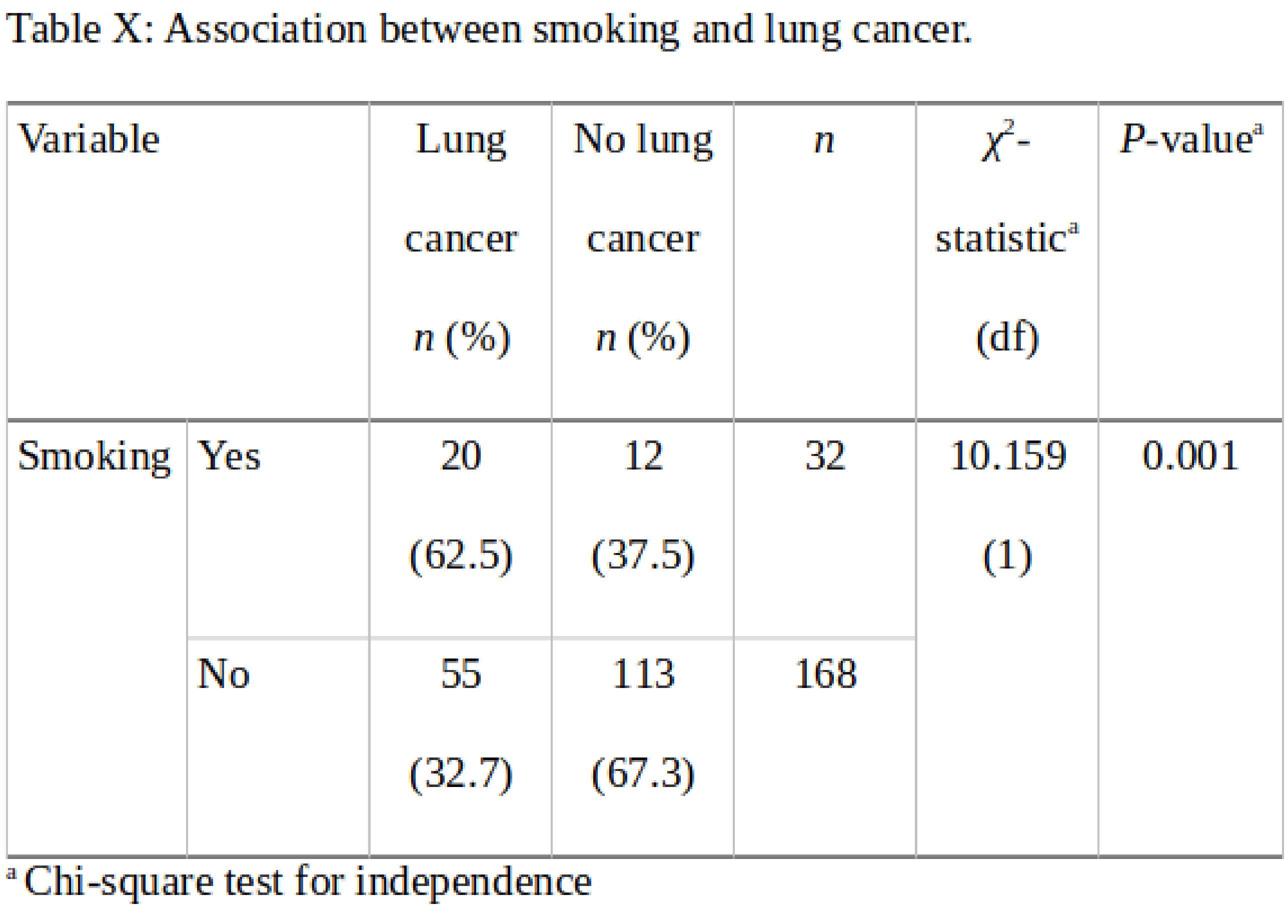


Table X: Association between smoking and lung cancer.

Variable		Lung cancer *n* (%)	No lung cancer *n* (%)	*n*	χ^2-statistic[a] (df)	*P*-value[a]
Smoking	Yes	20 (62.5)	12 (37.5)	32	10.159 (1)	0.001
	No	55 (32.7)	113 (67.3)	168		

[a] Chi-square test for independence

Fisher’s Exact Test

Fisher's Exact Test

- Purpose: Test the association between two categorical variables
- Situation:
    - When chi-squared test assumption not fulfilled
    - i.e. small expected count < 5 more 25% of the cells

Example

- Sample size: 20
- Variables:
  - Gender: Male / Female
  - Disease: Disease / No disease

Cross-tabulation

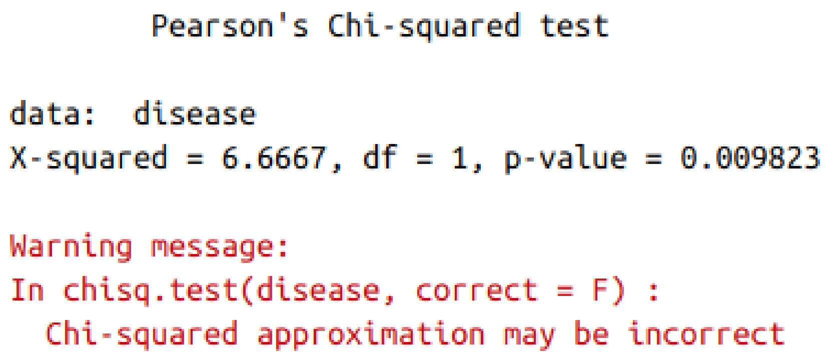
Gender	Disease	
	Disease	No disease
Male	10 (66.7%)	5 (33.3%)
Female	0 (0.0%)	5 (100.0%)

Expected Count

Gender	Disease		Sub-total
	Disease	No disease	
Male	10 (7.5)	5 (7.5)	15
Female	0 (2.5)	5 (2.5)	5
Sub-total	10	10	20

50% of expected count < 5, but none < 1

Results

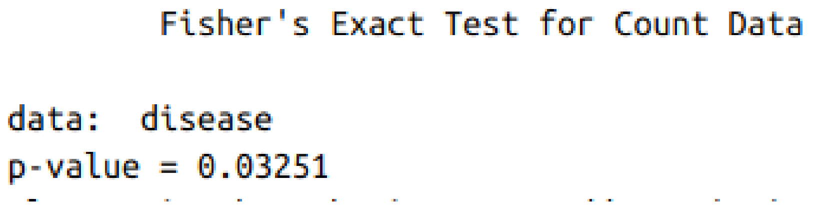


```
        Pearson's Chi-squared test

data:  disease
X-squared = 6.6667, df = 1, p-value = 0.009823

Warning message:
In chisq.test(disease, correct = F) :
  Chi-squared approximation may be incorrect
```

Using Chi-squared test is not appropriate



```
        Fisher's Exact Test for Count Data

data:  disease
p-value = 0.03251
```

Using Fisher’s exact

Results

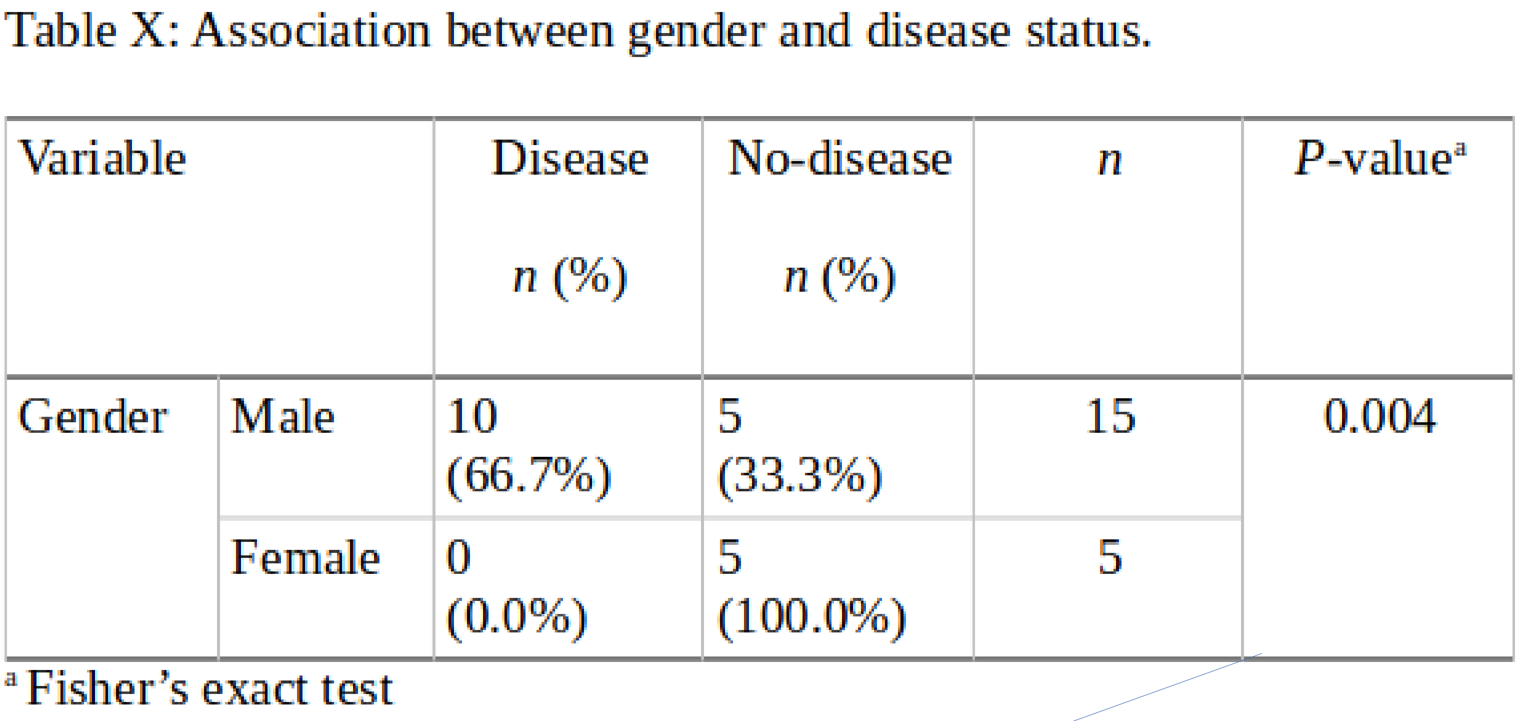


Table X: Association between gender and disease status.

Variable		Disease *n* (%)	No-disease *n* (%)	*n*	*P*-value[a]
Gender	Male	10 (66.7%)	5 (33.3%)	15	0.004
	Female	0 (0.0%)	5 (100.0%)	5	

[a] Fisher's exact test

No test statistic, only P-value

McNemar’s Test

McNemar’s Test

- Purpose: Test the difference between two repeated measurements of one categorical variable (two categories)
- e.g. pre-post treatment, paired measurement using different methods

McNemar's Test

- Whether the subjects still have the same outcomes (concordant) or different outcomes (discordant) upon repetition (pre-post)
- Determined by looking at the discordant cells
- Assumption:
  - Only two categories
  - Mutually exclusive categories

McNemar’s Test

Research objective:

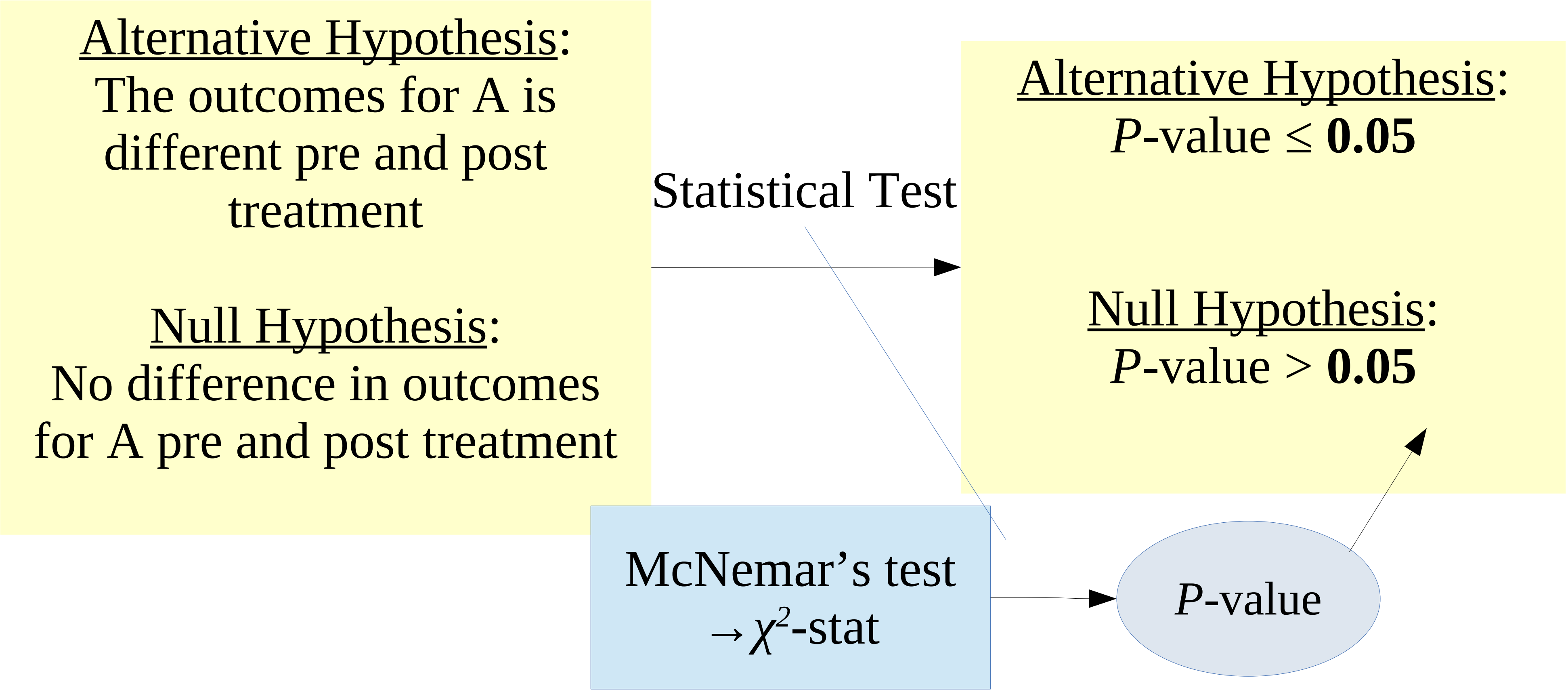
To test the difference in outcomes for A pre and post treatment

Research question:

Is there any difference outcomes for A pre and post treatment?

McNemar’s Test

RQ: Is there any difference in outcomes for A pre and post treatment?



Alternative Hypothesis:
The outcomes for A is different pre and post treatment

Null Hypothesis:
No difference in outcomes for A pre and post treatment

Statistical Test

Alternative Hypothesis:
P-value ≤ **0.05**

Null Hypothesis:
P-value > **0.05**

McNemar’s test
→ χ^2-stat

P-value

Example

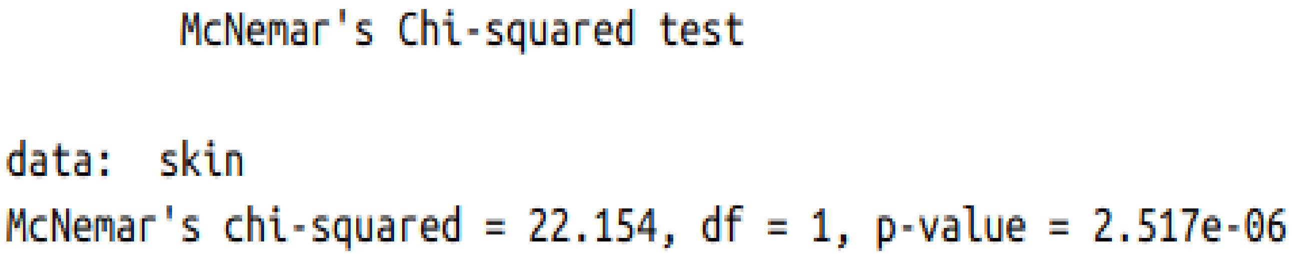
- Sample size: 60
- Variable:
  - Size of skin lesion pre and post treatment

Cross-tabulation

Skin Lesion Size Before Treatment	Skin Lesion Size After Treatment		Sub-total
	Large	Small	
Large	5	25	30
Small	1	29	30
Sub-total	6	54	60

Discordant pairs

Results



```
        McNemar's Chi-squared test

data:  skin
McNemar's chi-squared = 22.154, df = 1, p-value = 2.517e-06
```

McNemar’s test uses chi-squared statistics to get *P*-value

Results

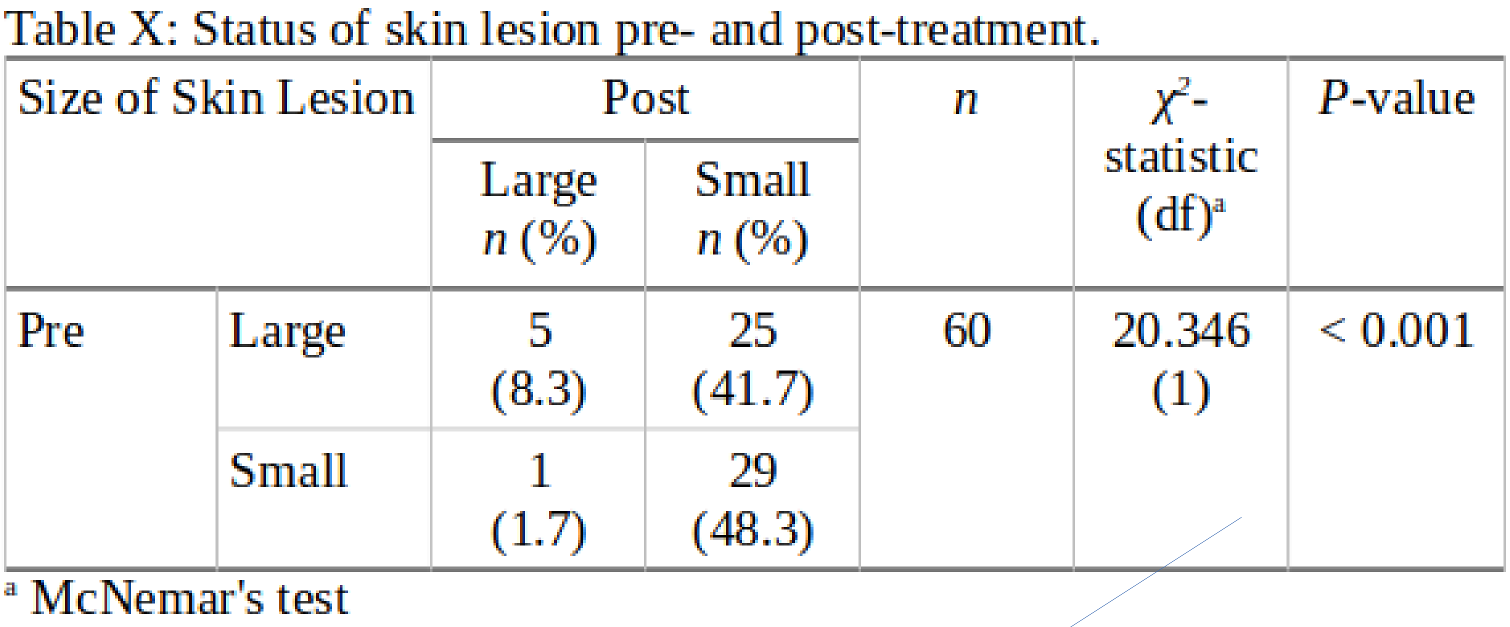


Table X: Status of skin lesion pre- and post-treatment.

Size of Skin Lesion		Post		*n*	χ^2-statistic (df)[a]	*P*-value
		Large *n* (%)	Small *n* (%)			
Pre	Large	5 (8.3)	25 (41.7)	60	20.346 (1)	< 0.001
	Small	1 (1.7)	29 (48.3)			

[a] McNemar's test

McNemar's test also uses X^2 statistics

Quiz

- Briefly describe about parametric test
- Describe the purpose of testing by Chi-squared test
- Describe the purpose of testing by Fisher’s exact test
- Describe the purpose of testing by McNemar’s test

Quiz

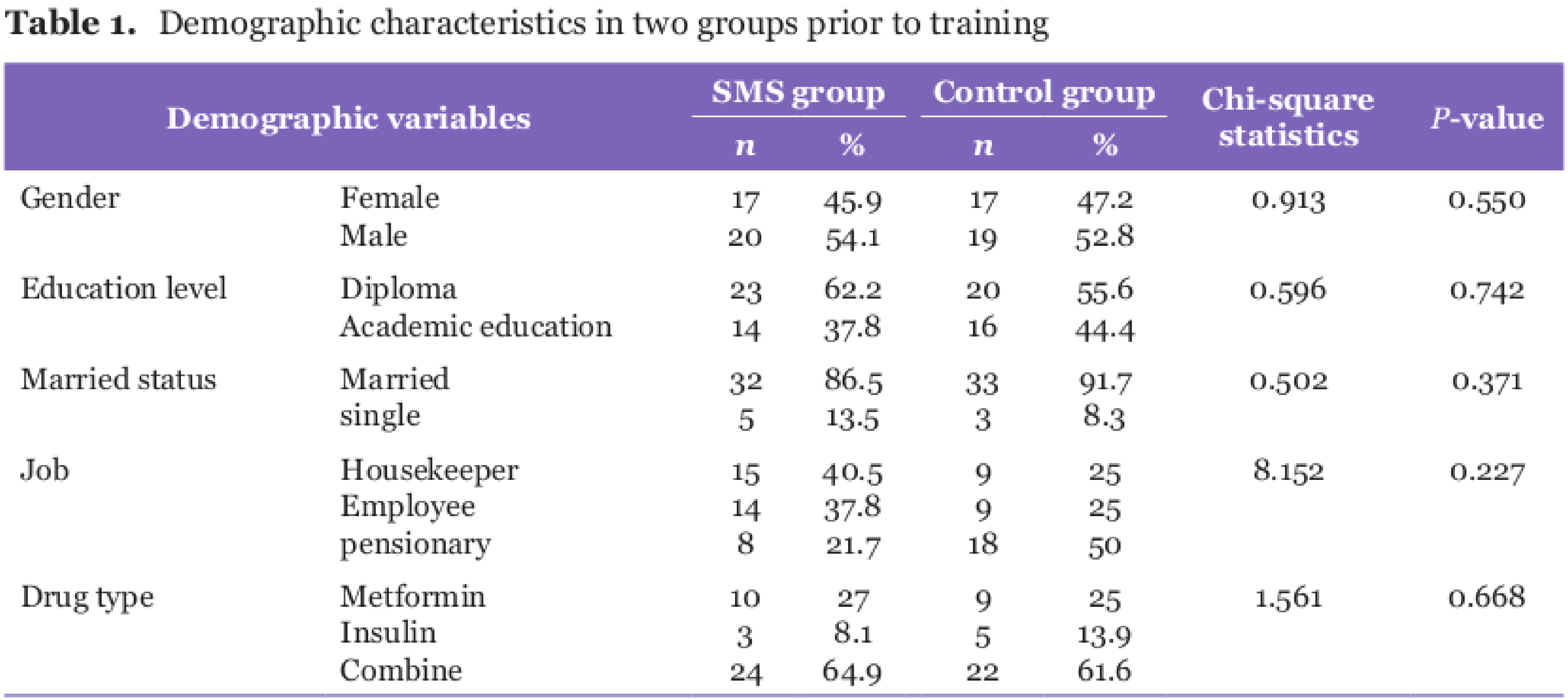


Table 1. Demographic characteristics in two groups prior to training

Demographic variables		SMS group		Control group		Chi-square statistics	*P*-value
		n	%	*n*	%		
Gender	Female	17	45.9	17	47.2	0.913	0.550
	Male	20	54.1	19	52.8		
Education level	Diploma	23	62.2	20	55.6	0.596	0.742
	Academic education	14	37.8	16	44.4		
Married status	Married	32	86.5	33	91.7	0.502	0.371
	single	5	13.5	3	8.3		
Job	Housekeeper	15	40.5	9	25	8.152	0.227
	Employee	14	37.8	9	25		
	pensionary	8	21.7	18	50		
Drug type	Metformin	10	27	9	25	1.561	0.668
	Insulin	3	8.1	5	13.9		
	Combine	24	64.9	22	61.6		

Lari, H., Noroozi, A., & Tahmasebi, R. (2018). Impact of short message service (SMS) education based on a health promotion model on the physical activity of patients with type II diabetes. The Malaysian journal of medical sciences: MJMS, 25(3), 67.

Quiz

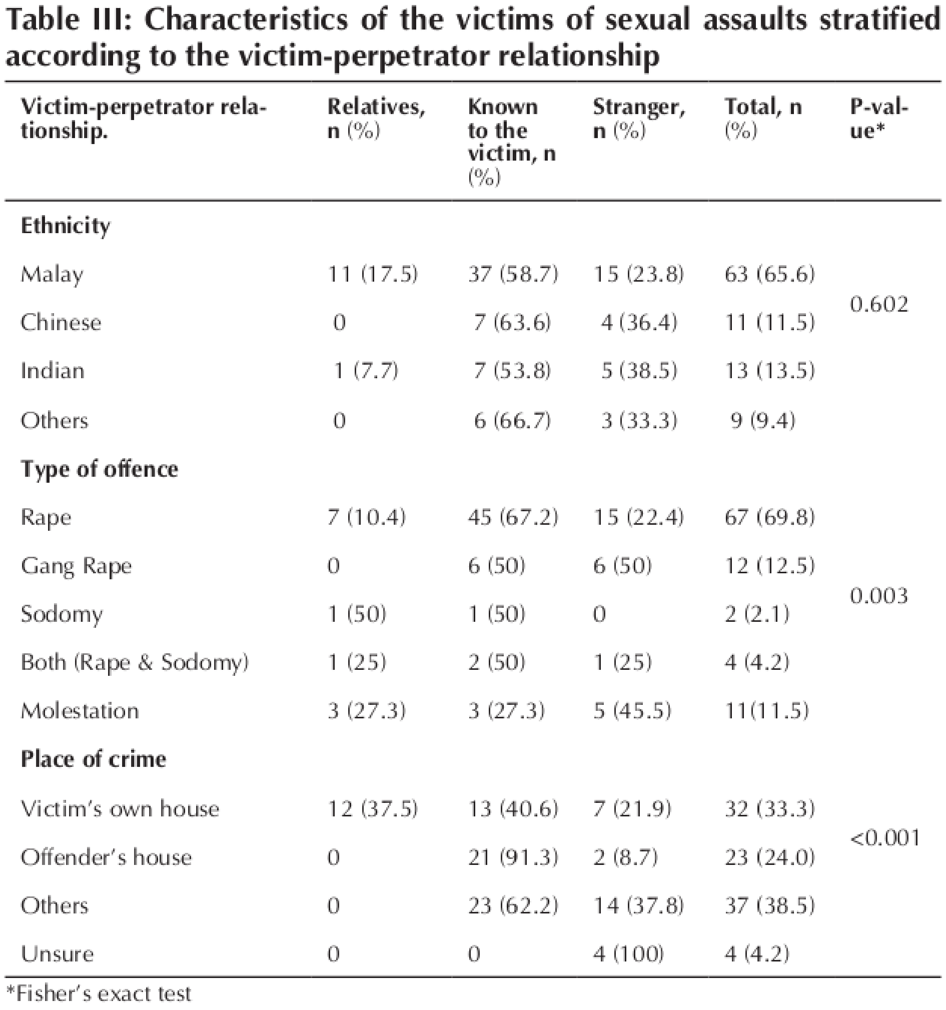


Table III: Characteristics of the victims of sexual assaults stratified according to the victim-perpetrator relationship

Victim-perpetrator relationship.	Relatives, n (%)	Known to the victim, n (%)	Stranger, n (%)	Total, n (%)	P-value*
Ethnicity					
Malay	11 (17.5)	37 (58.7)	15 (23.8)	63 (65.6)	0.602
Chinese	0	7 (63.6)	4 (36.4)	11 (11.5)	
Indian	1 (7.7)	7 (53.8)	5 (38.5)	13 (13.5)	
Others	0	6 (66.7)	3 (33.3)	9 (9.4)	
Type of offence					
Rape	7 (10.4)	45 (67.2)	15 (22.4)	67 (69.8)	0.003
Gang Rape	0	6 (50)	6 (50)	12 (12.5)	
Sodomy	1 (50)	1 (50)	0	2 (2.1)	
Both (Rape & Sodomy)	1 (25)	2 (50)	1 (25)	4 (4.2)	
Molestation	3 (27.3)	3 (27.3)	5 (45.5)	11(11.5)	
Place of crime					
Victim's own house	12 (37.5)	13 (40.6)	7 (21.9)	32 (33.3)	<0.001
Offender's house	0	21 (91.3)	2 (8.7)	23 (24.0)	
Others	0	23 (62.2)	14 (37.8)	37 (38.5)	
Unsure	0	0	4 (100)	4 (4.2)	

*Fisher's exact test

Ahmad, M. I., Ismail, R., Arifin, W. N., Noordin, M., Amirah, N., Bahari, N. S. N. S., & Arshad, M. K. N. M. (2020). Sexual Assault: A Descriptive Study of Victims Attending a Public Hospital in Ipoh. Malaysian Journal of Medicine & Health Sciences, 16(1).

Quiz

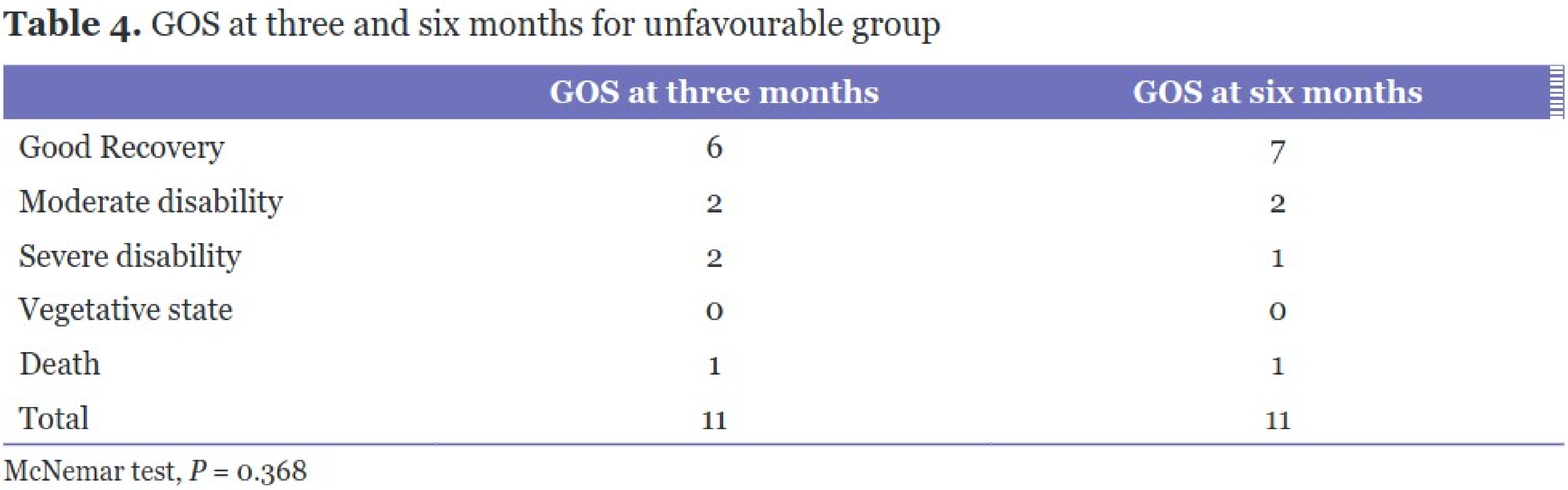


Table 4. GOS at three and six months for unfavourable group

	GOS at three months	GOS at six months
Good Recovery	6	7
Moderate disability	2	2
Severe disability	2	1
Vegetative state	0	0
Death	1	1
Total	11	11

McNemar test, $P = 0.368$

Sidek, M. S. M., Siregar, J. A., Ghani, A. R. I., & Idris, Z. (2018). Teleneurosurgery: outcome of mild head injury patients managed in non-neurosurgical centre in the state of Johor. The Malaysian journal of medical sciences: MJMS, 25(2), 95.

Thank You