Prism Manual
Categorical Data Analysis

Jingwen Gu
BCBB/OCICB/NIAID

This lab will introduce categorical data analysis using Prism. For more analysis for categorical data analysis in Prism, please check chapter in Prism statistics guide. To learn more about categorical data analysis, please check another BCBB workshop – Categorical data analysis.

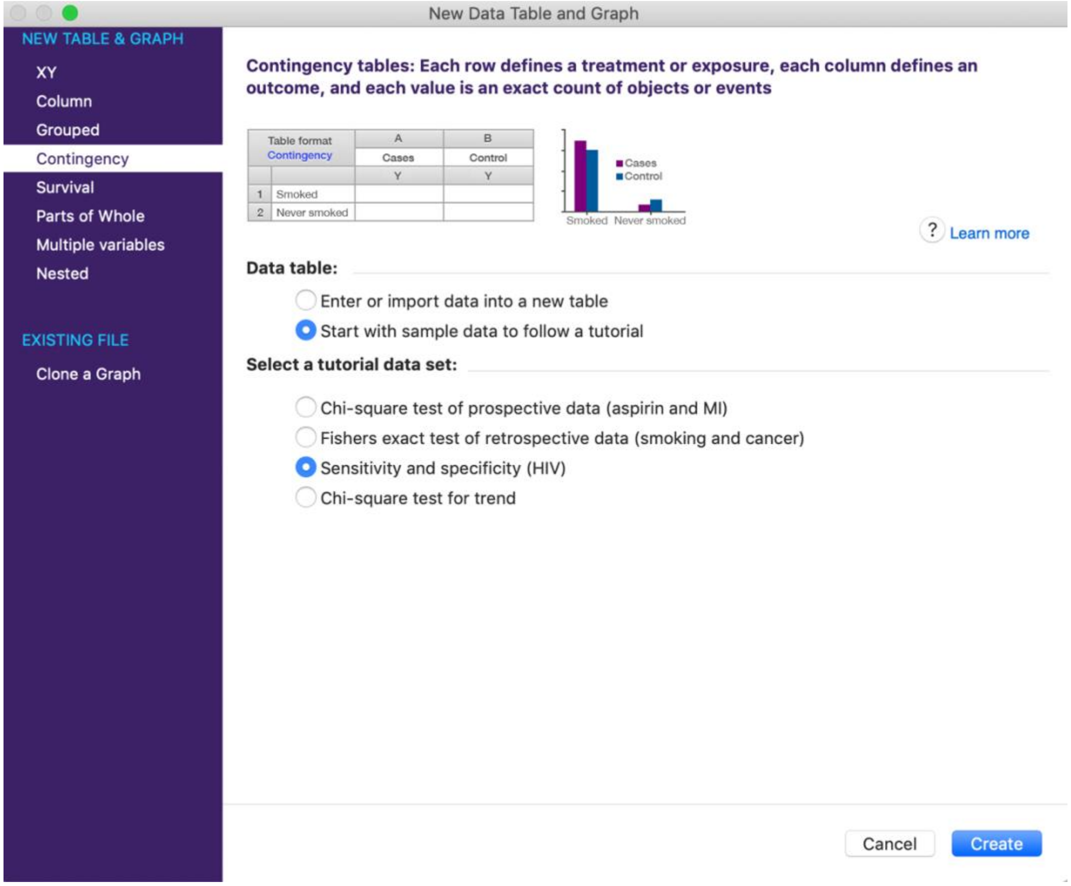
0. Outline:

1) Create a contingency table
2) Calculation - relative risk, difference between proportion, odds ratio, sensitivity etc (All of these applies to 2 × 2 tables)
3) Perform statistical testing for contingency table

1. Create a contingency table

You must enter data in the form of a contingency table. Prism cannot cross-tabulate raw data to create a contingency table.

1.1 Select **Contingency** under "NEW TABLE & GRAPH" menu. Then select "**Start with sample data to follow a tutorial**" or the first option to enter your data. Select "**Sensitivity and specificity (HIV)**". Click create.



Contingency tables: Each row defines a treatment or exposure, each column defines an outcome, and each value is an exact count of objects or events

Data table:

○ Enter or import data into a new table

● Start with sample data to follow a tutorial

Select a tutorial data set:

○ Chi-square test of prospective data (aspirin and MI)

○ Fishers exact test of retrospective data (smoking and cancer)

● Sensitivity and specificity (HIV)

○ Chi-square test for trend

2. Calculation

2.1 Click **Analyze**

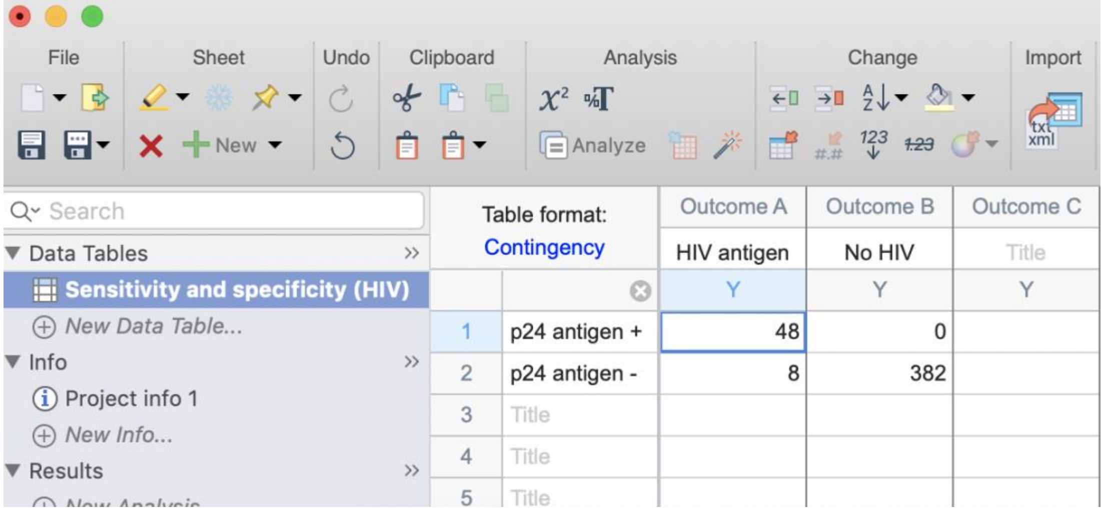
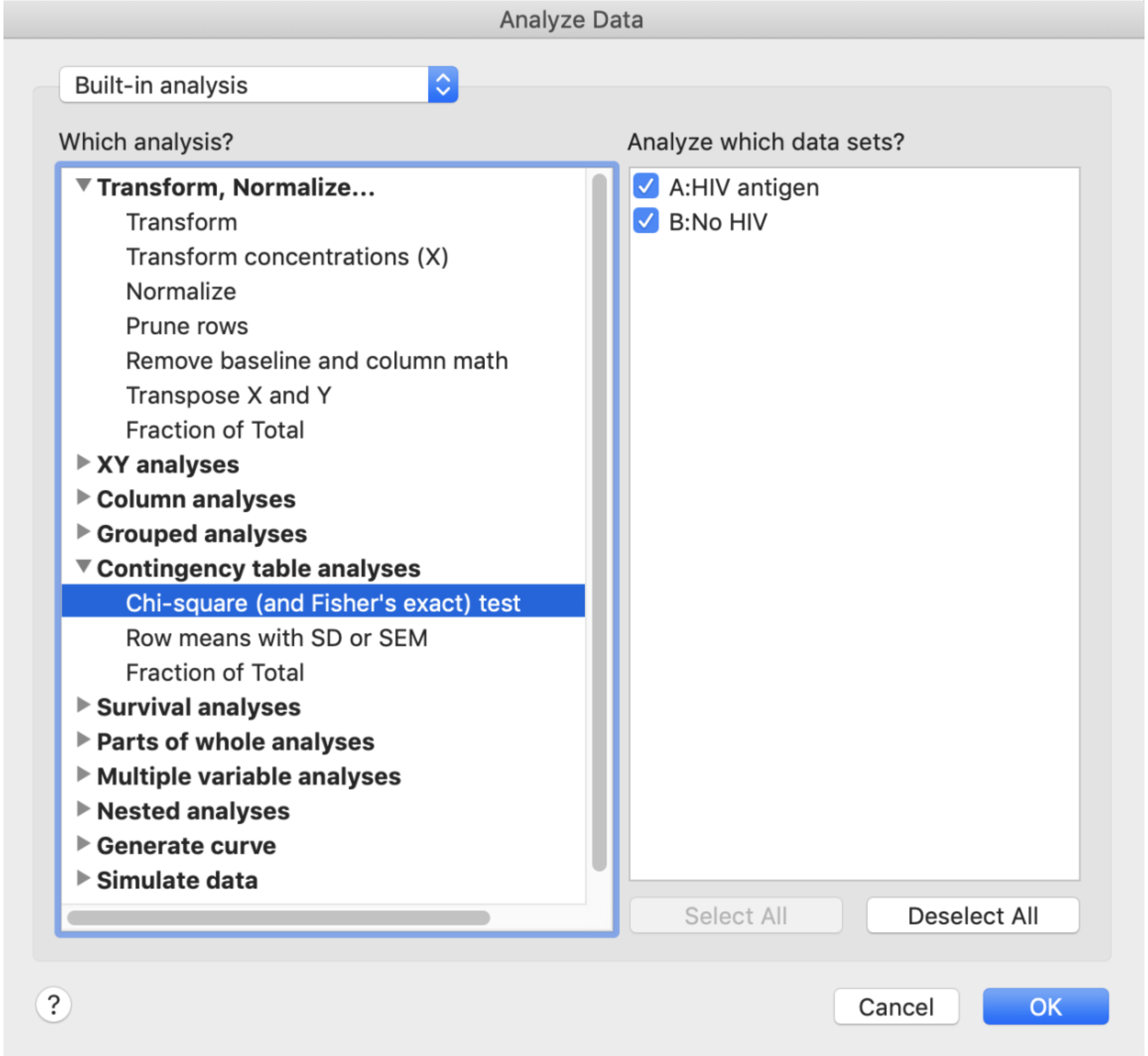


	Table format: Contingency	Outcome A	Outcome B	Outcome C
		HIV antigen	No HIV	Title
		Y	Y	Y
1	p24 antigen +	48	0	
2	p24 antigen -	8	382	
3	Title			
4	Title			
5	Title			

2.2 Select **Chi-square (and Fisher's exact) test**, then OK.



Built-in analysis

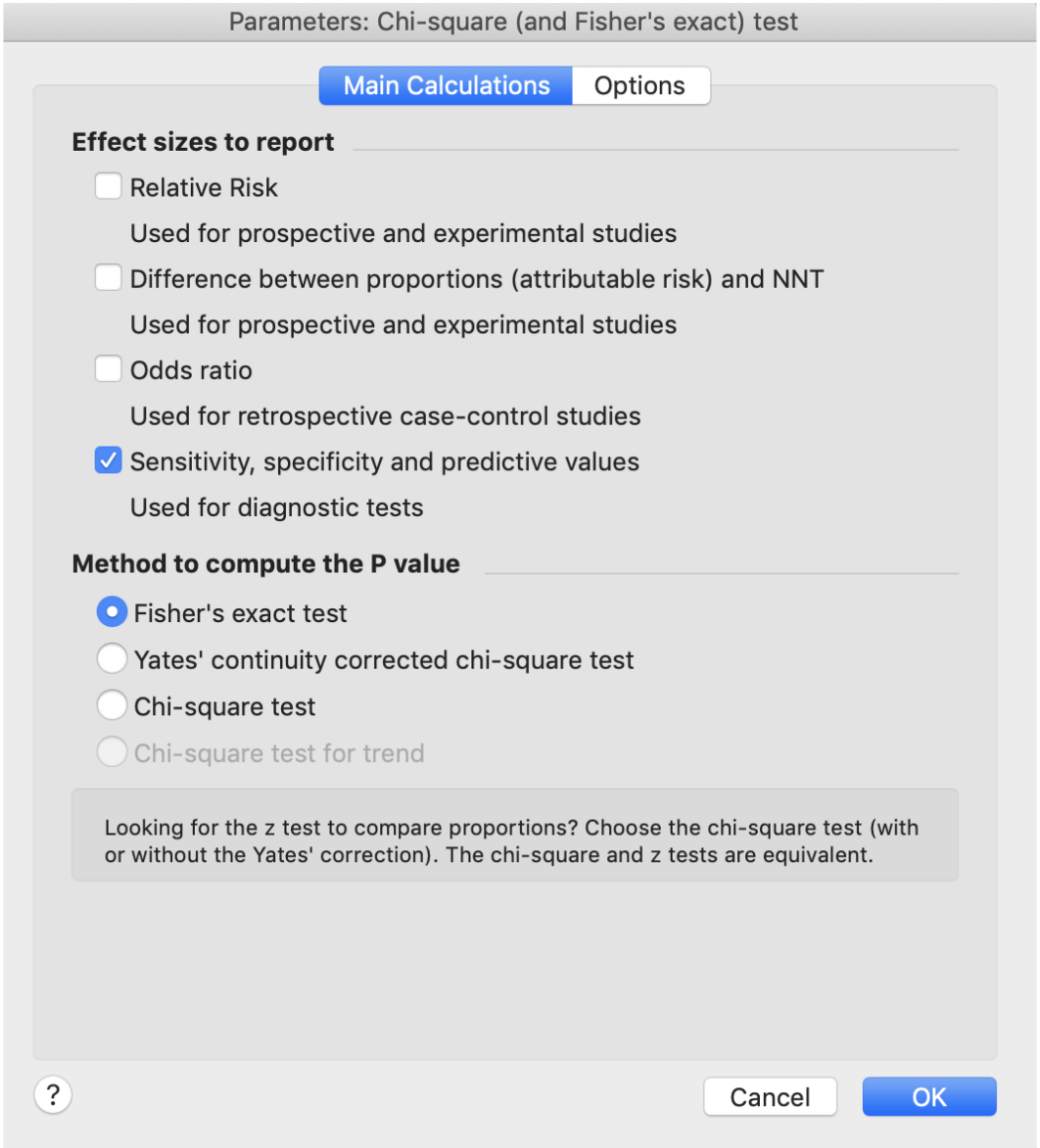
Which analysis? Chi-square (and Fisher's exact) test

Analyze which data sets?

☑ A:HIV antigen

☑ B:No HIV

2.3 Select the **type of calculation** you want under Main Calculations. You could decide one sided or two sided from Options. Here we select **Sensitivity, specificity and predictive values** as an example.



Parameters: Chi-square (and Fisher's exact) test

Effect sizes to report

□ Relative Risk

Used for prospective and experimental studies

□ Difference between proportions (attributable risk) and NNT

Used for prospective and experimental studies

□ Odds ratio

Used for retrospective case-control studies

☑ Sensitivity, specificity and predictive values

Used for diagnostic tests

Method to compute the P value

● Fisher's exact test

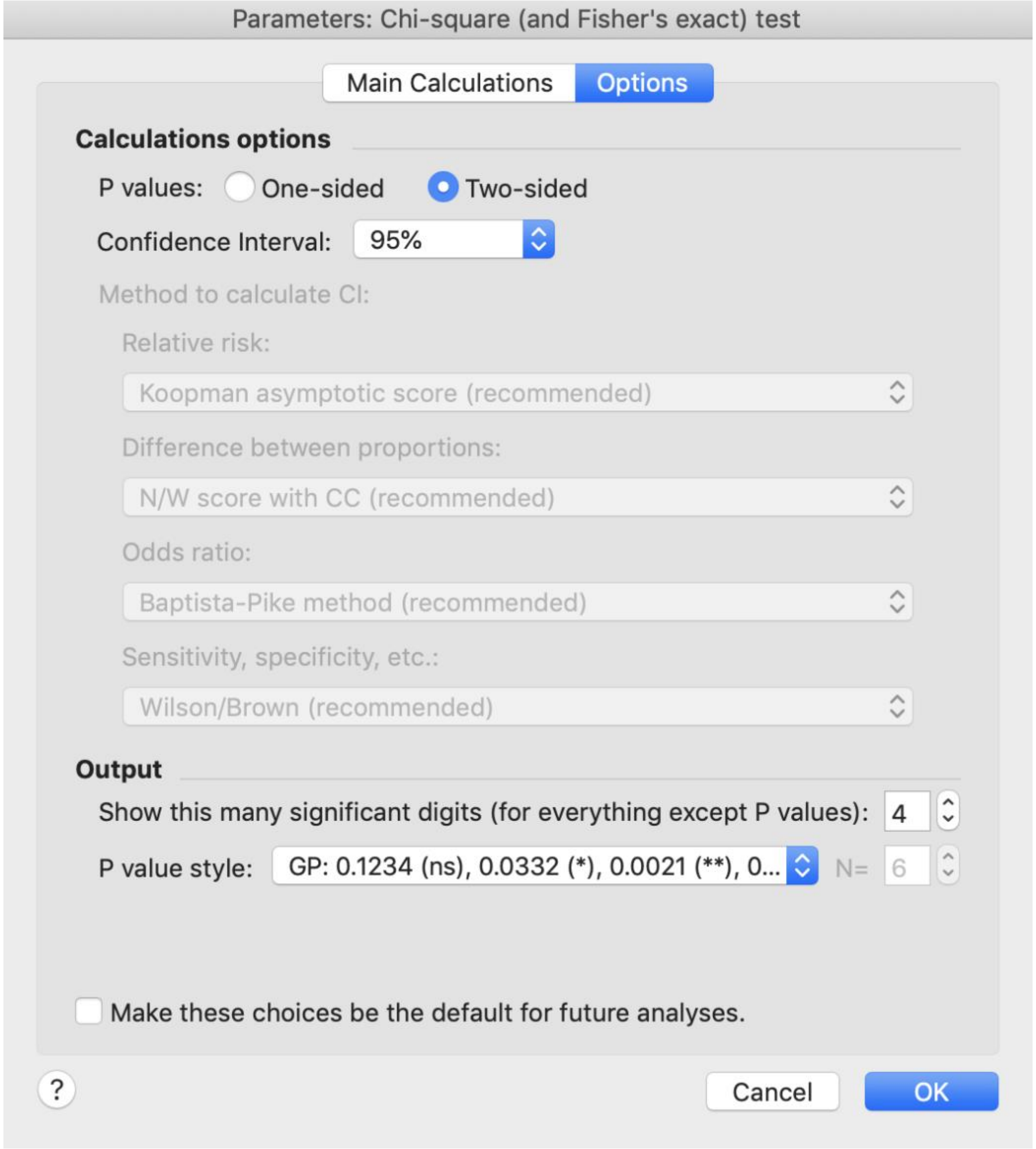
○ Yates' continuity corrected chi-square test

○ Chi-square test

○ Chi-square test for trend

Looking for the z test to compare proportions? Choose the chi-square test (with or without the Yates' correction). The chi-square and z tests are equivalent.

A two-sided P value is recommended unless you have a strong reason to choose a one-sided P value.



Parameters: Chi-square (and Fisher's exact) test

Calculations options

P values: ○ One-sided ● Two-sided

Confidence Interval: 95%

Method to calculate CI:

Relative risk: Koopman asymptotic score (recommended)

Difference between proportions: N/W score with CC (recommended)

Odds ratio: Baptista-Pike method (recommended)

Sensitivity, specificity, etc.: Wilson/Brown (recommended)

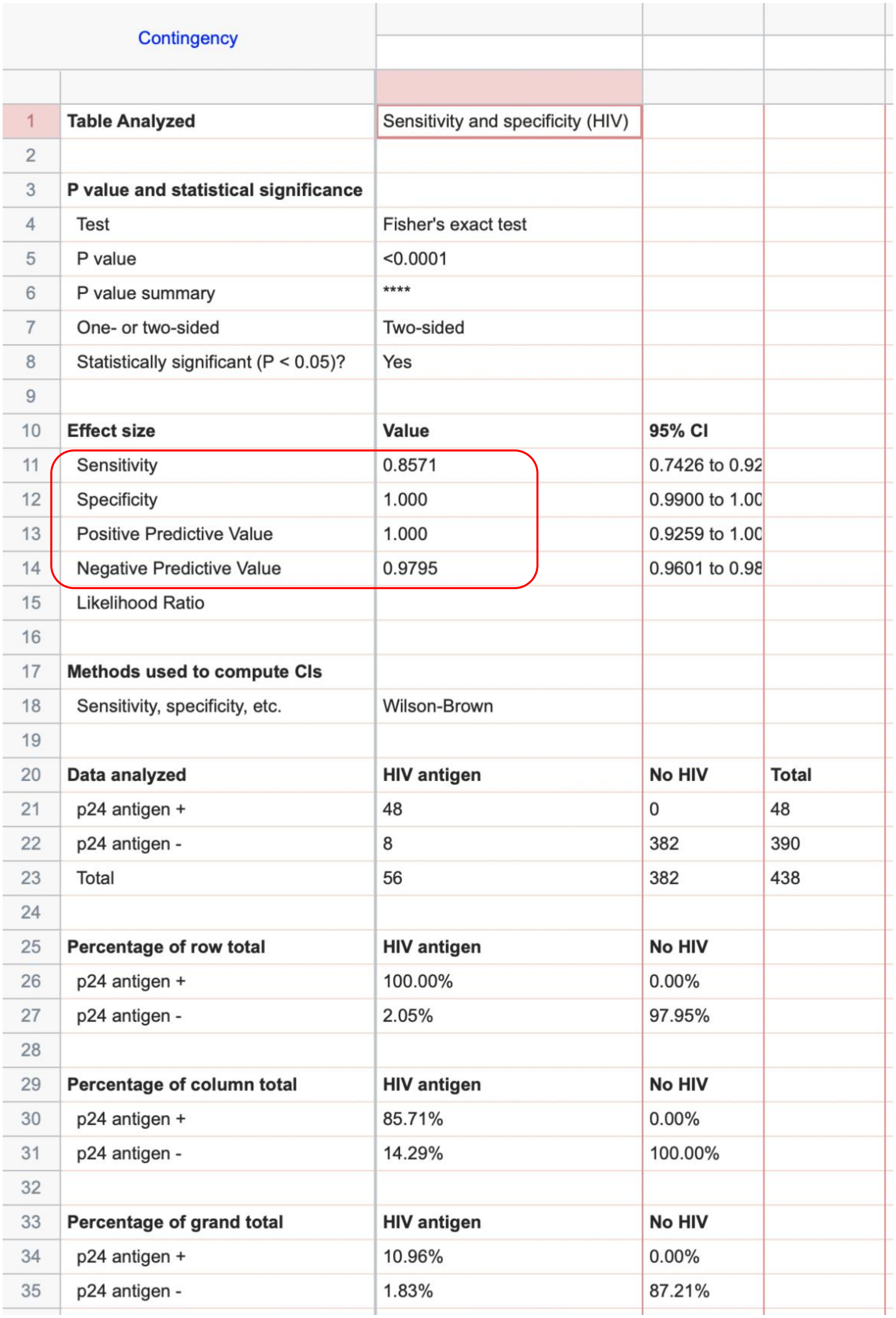
Output

Show this many significant digits (for everything except P values): 4

P value style: GP: 0.1234 (ns), 0.0332 (*), 0.0021 (**), 0... N= 6

□ Make these choices be the default for future analyses.

2.4 Result are circled. More details about reporting the result could be found at here.



Contingency

Table Analyzed	Sensitivity and specificity (HIV)		
P value and statistical significance			
Test	Fisher's exact test		
P value	<0.0001		
P value summary	****		
One- or two-sided	Two-sided		
Statistically significant (P < 0.05)?	Yes		
Effect size	**Value**	**95% CI**	
Sensitivity	0.8571	0.7426 to 0.92	
Specificity	1.000	0.9900 to 1.00	
Positive Predictive Value	1.000	0.9259 to 1.00	
Negative Predictive Value	0.9795	0.9601 to 0.98	
Likelihood Ratio			
Methods used to compute CIs			
Sensitivity, specificity, etc.	Wilson-Brown		
Data analyzed	**HIV antigen**	**No HIV**	**Total**
p24 antigen +	48	0	48
p24 antigen -	8	382	390
Total	56	382	438
Percentage of row total	**HIV antigen**	**No HIV**	
p24 antigen +	100.00%	0.00%	
p24 antigen -	2.05%	97.95%	
Percentage of column total	**HIV antigen**	**No HIV**	
p24 antigen +	85.71%	0.00%	
p24 antigen -	14.29%	100.00%	
Percentage of grand total	**HIV antigen**	**No HIV**	
p24 antigen +	10.96%	0.00%	
p24 antigen -	1.83%	87.21%	

When testing the independence of categorical data, different tests could be selected based on data characteristic:

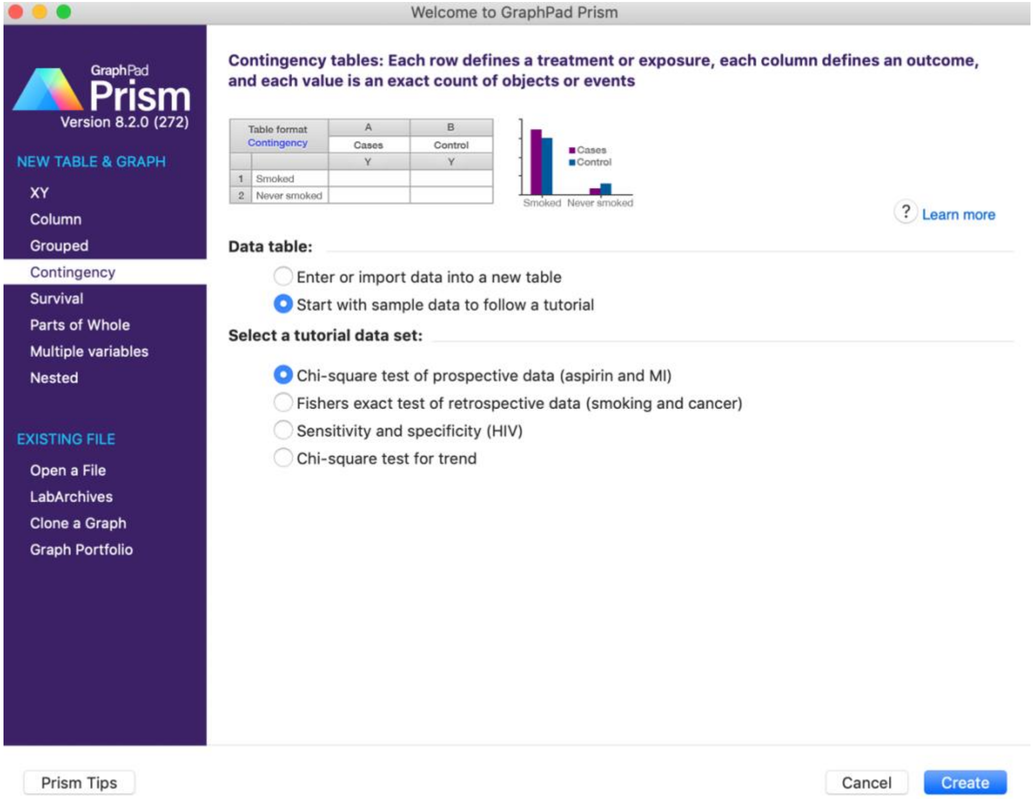
For example, if we have data with large sample size and independence, Pearson chi-square test and likelihood ratio test can be used. Fisher's exact test uses when sample size of data is small and conditioned. Cochran-Mantel-Haenszel test uses with stratified data. McNemar's test or CMH test uses with paired data. Linear trend test with ordinal data.

Among these, Chi-square test and Fisher's exact test are available in Prism software and McNemar test is available in the Prism online calculator. The null hypothesis of the test is no association, alternative hypothesis is there is association.

3.1 Pearson Chi-square test

The chi-squared test is used to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more categories.

3.1.1 Select **Contingency** under "New Table & Graph"; Select **Start with sample data to follow a tutorial** and **Chi-square test of prospective data (aspirin and MI)**, then click **Create**.



Welcome to GraphPad Prism

GraphPad Prism
Version 8.2.0 (272)

NEW TABLE & GRAPH
- XY
- Column
- Grouped
- Contingency
- Survival
- Parts of Whole
- Multiple variables
- Nested

EXISTING FILE
- Open a File
- LabArchives
- Clone a Graph
- Graph Portfolio

Contingency tables: Each row defines a treatment or exposure, each column defines an outcome, and each value is an exact count of objects or events

Table format Contingency	A Cases Y	B Control Y
1 Smoked		
2 Never smoked		

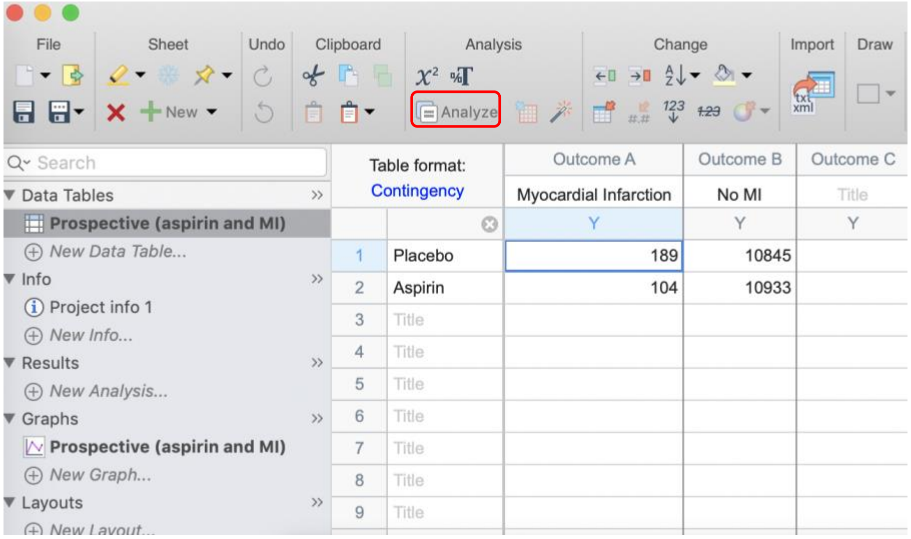
Data table:

○ Enter or import data into a new table
● Start with sample data to follow a tutorial

Select a tutorial data set:

● Chi-square test of prospective data (aspirin and MI)
○ Fishers exact test of retrospective data (smoking and cancer)
○ Sensitivity and specificity (HIV)
○ Chi-square test for trend

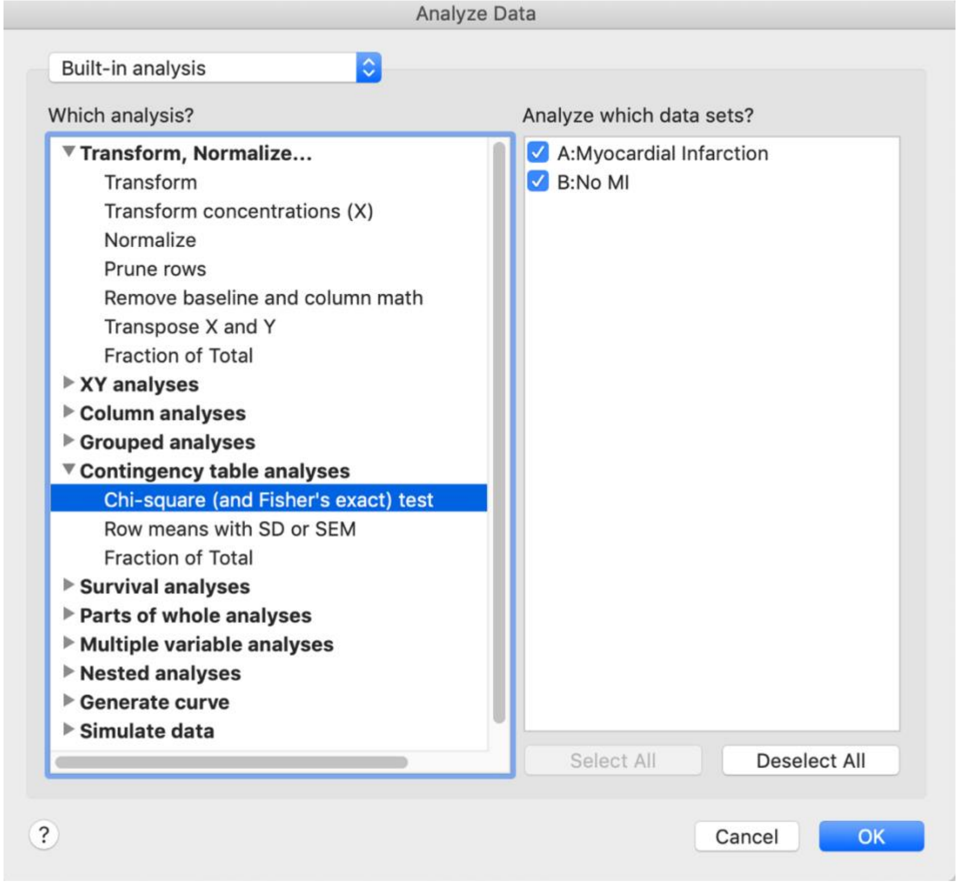
3.1.2 Click **Analyze** Analyze after sample data opened.



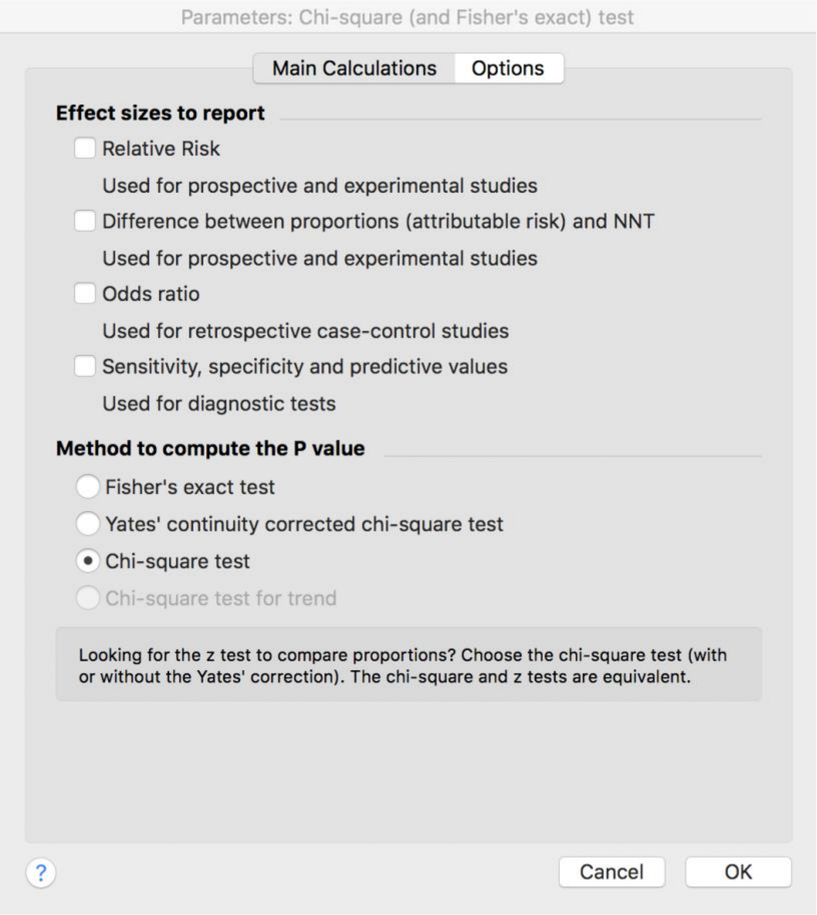
- Data Tables
  - Prospective (aspirin and MI)
  - New Data Table...
- Info
  - Project info 1
  - New Info...
- Results
  - New Analysis...
- Graphs
  - Prospective (aspirin and MI)
  - New Graph...
- Layouts
  - New Layout...

Table format: Contingency		Outcome A	Outcome B	Outcome C
		Myocardial Infarction	No MI	Title
		Y	Y	Y
1	Placebo	189	10845	
2	Aspirin	104	10933	
3	Title			
4	Title			
5	Title			
6	Title			
7	Title			
8	Title			
9	Title			

3.1.3 Select **Chi-square (and Fisher's exact) test** under Contingency table analyses list. Check the columns you want to analyze on the right side, then click **OK**.



3.1.4 Select test type under Method to compute P value and other report statistic under Effect sizes to report. Click **Chi-square test**, then **OK**.



Parameters: Chi-square (and Fisher's exact) test

Effect sizes to report

□ Relative Risk

Used for prospective and experimental studies

□ Difference between proportions (attributable risk) and NNT

Used for prospective and experimental studies

□ Odds ratio

Used for retrospective case-control studies

□ Sensitivity, specificity and predictive values

Used for diagnostic tests

Method to compute the P value

○ Fisher's exact test

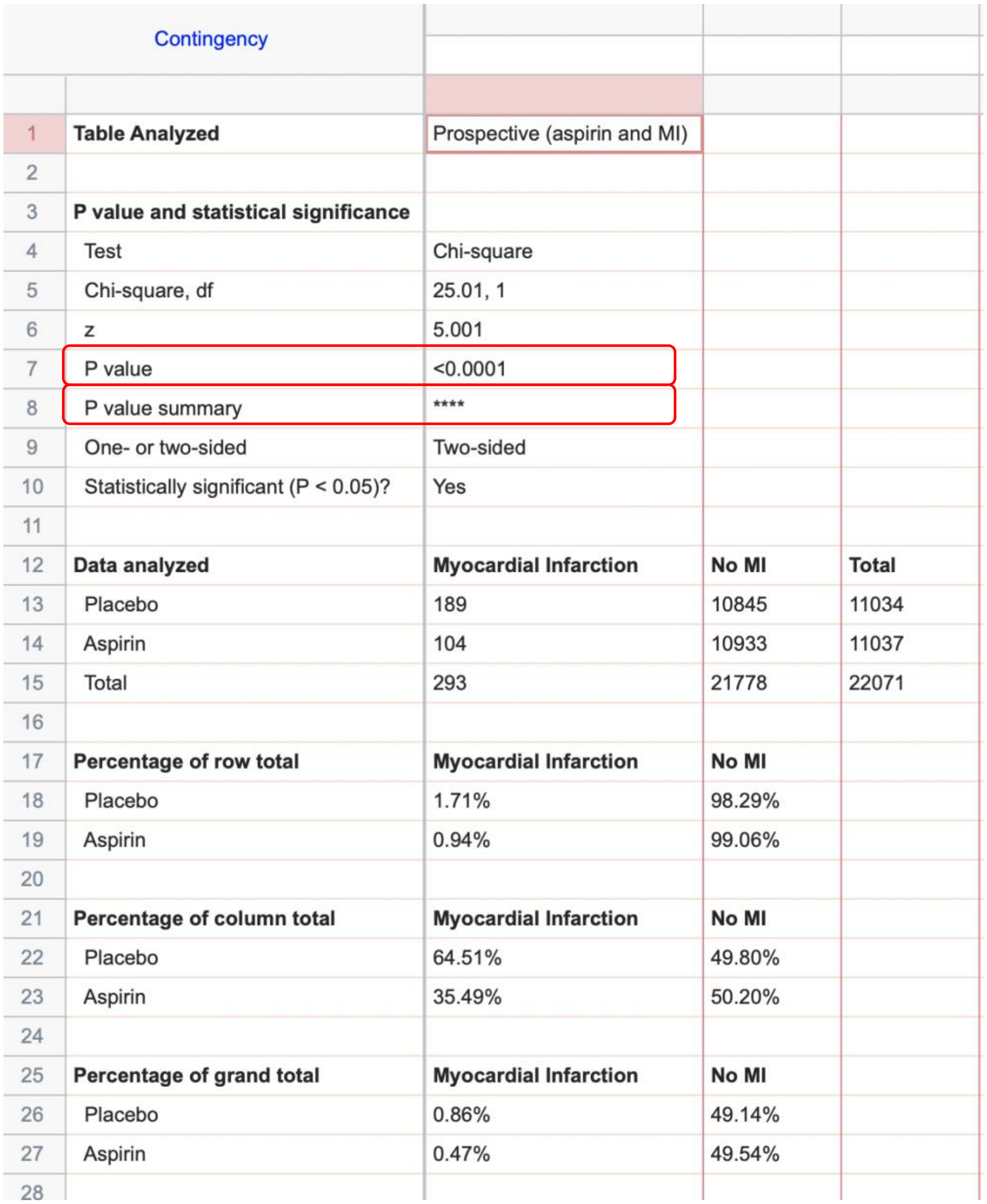
○ Yates' continuity corrected chi-square test

● Chi-square test

○ Chi-square test for trend

Looking for the z test to compare proportions? Choose the chi-square test (with or without the Yates' correction). The chi-square and z tests are equivalent.

3.1.5 Result includes p-value, marginal table, and marginal percentage. Chi-square statistic, degree of freedom and p-value are circled. Test result is significant, which means we have 95% confidence to reject the null hypothesis and conclude that there is association between aspirin and MI.



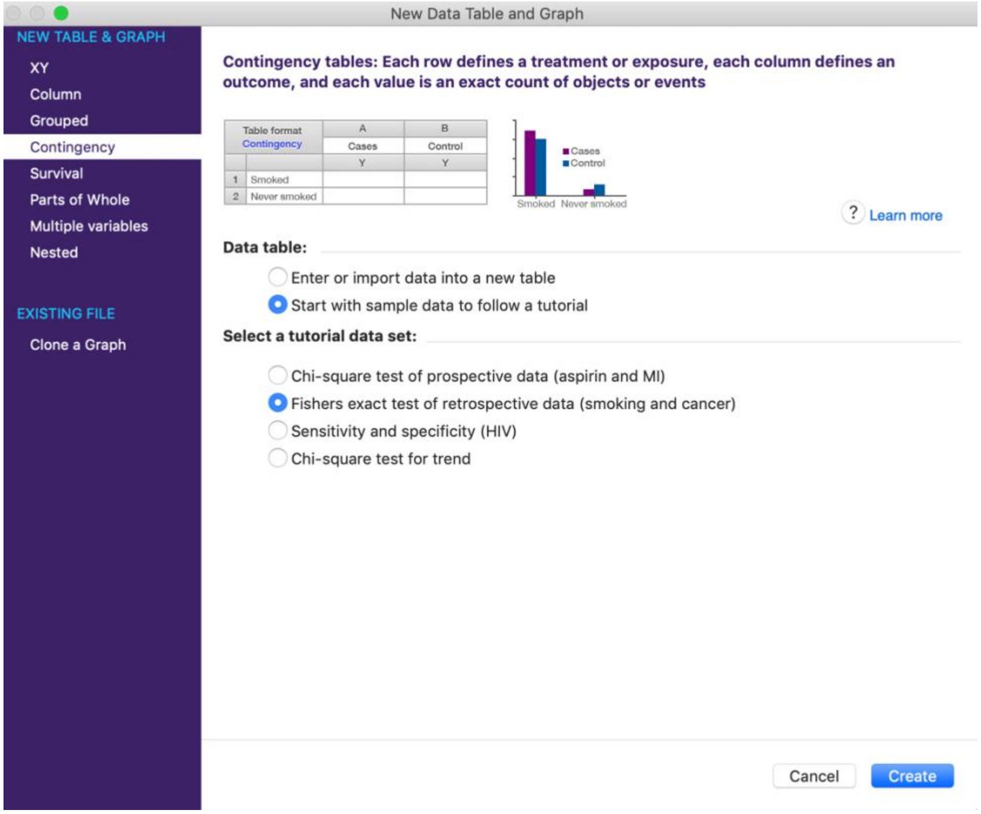
Contingency

Table Analyzed	Prospective (aspirin and MI)		
P value and statistical significance			
Test	Chi-square		
Chi-square, df	25.01, 1		
z	5.001		
P value	<0.0001		
P value summary	****		
One- or two-sided	Two-sided		
Statistically significant (P < 0.05)?	Yes		
Data analyzed	**Myocardial Infarction**	**No MI**	**Total**
Placebo	189	10845	11034
Aspirin	104	10933	11037
Total	293	21778	22071
Percentage of row total	**Myocardial Infarction**	**No MI**	
Placebo	1.71%	98.29%	
Aspirin	0.94%	99.06%	
Percentage of column total	**Myocardial Infarction**	**No MI**	
Placebo	64.51%	49.80%	
Aspirin	35.49%	50.20%	
Percentage of grand total	**Myocardial Infarction**	**No MI**	
Placebo	0.86%	49.14%	
Aspirin	0.47%	49.54%	

3.2 Fisher's exact test

Fisher's exact test is optimal to test association between small and conditioned sample size. It is named after its inventor, Ronald Fisher. It is an exact test because the significance of the deviation from a null hypothesis (e.g., P-value) can be calculated exactly, rather than relying on an approximation that becomes exact in the limit as the sample size grows to infinity.

3.2.1 Select **Contingency** under "New Table & Graph"; Select **Start with sample data to follow a tutorial** and **Fishers exact test of retrospective data (smoking and cancer)**, then click **Create**.



New Data Table and Graph

NEW TABLE & GRAPH
- XY
- Column
- Grouped
- Contingency
- Survival
- Parts of Whole
- Multiple variables
- Nested

EXISTING FILE
- Clone a Graph

Contingency tables: Each row defines a treatment or exposure, each column defines an outcome, and each value is an exact count of objects or events

Table format Contingency	A Cases	B Control
	Y	Y
1 Smoked		
2 Never smoked		

? Learn more

Data table:

○ Enter or import data into a new table

● Start with sample data to follow a tutorial

Select a tutorial data set:

○ Chi-square test of prospective data (aspirin and MI)

● Fishers exact test of retrospective data (smoking and cancer)

○ Sensitivity and specificity (HIV)

○ Chi-square test for trend

Cancel | Create

3.2.2 Click **Analyze** Analyze after sample data opened.

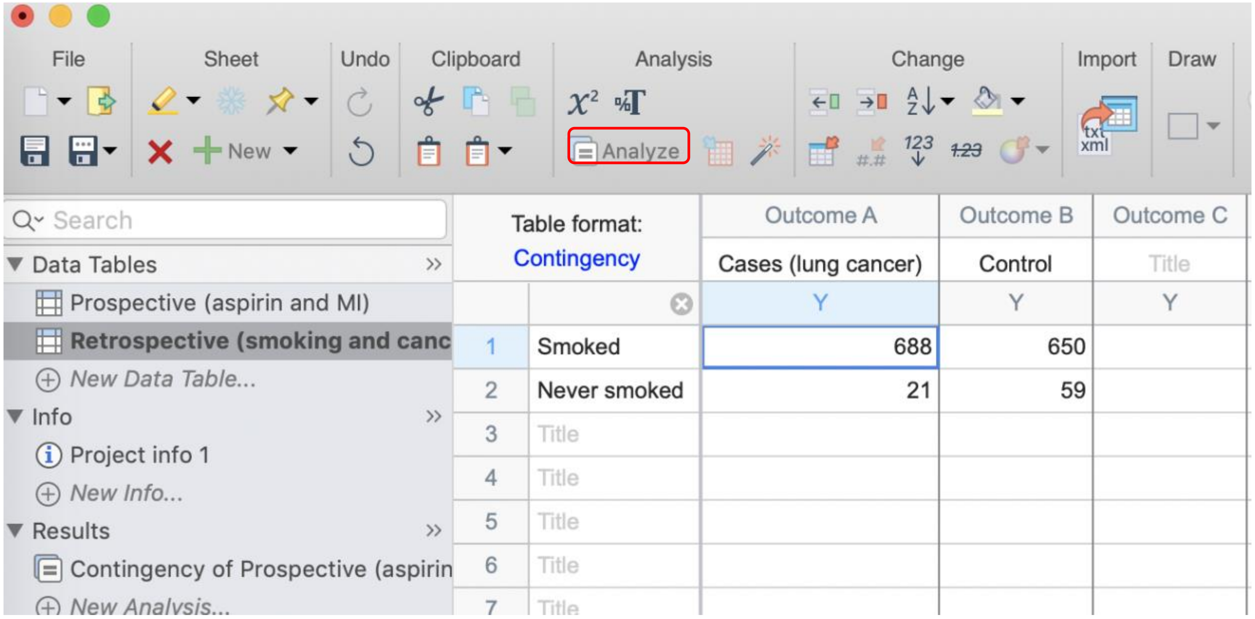
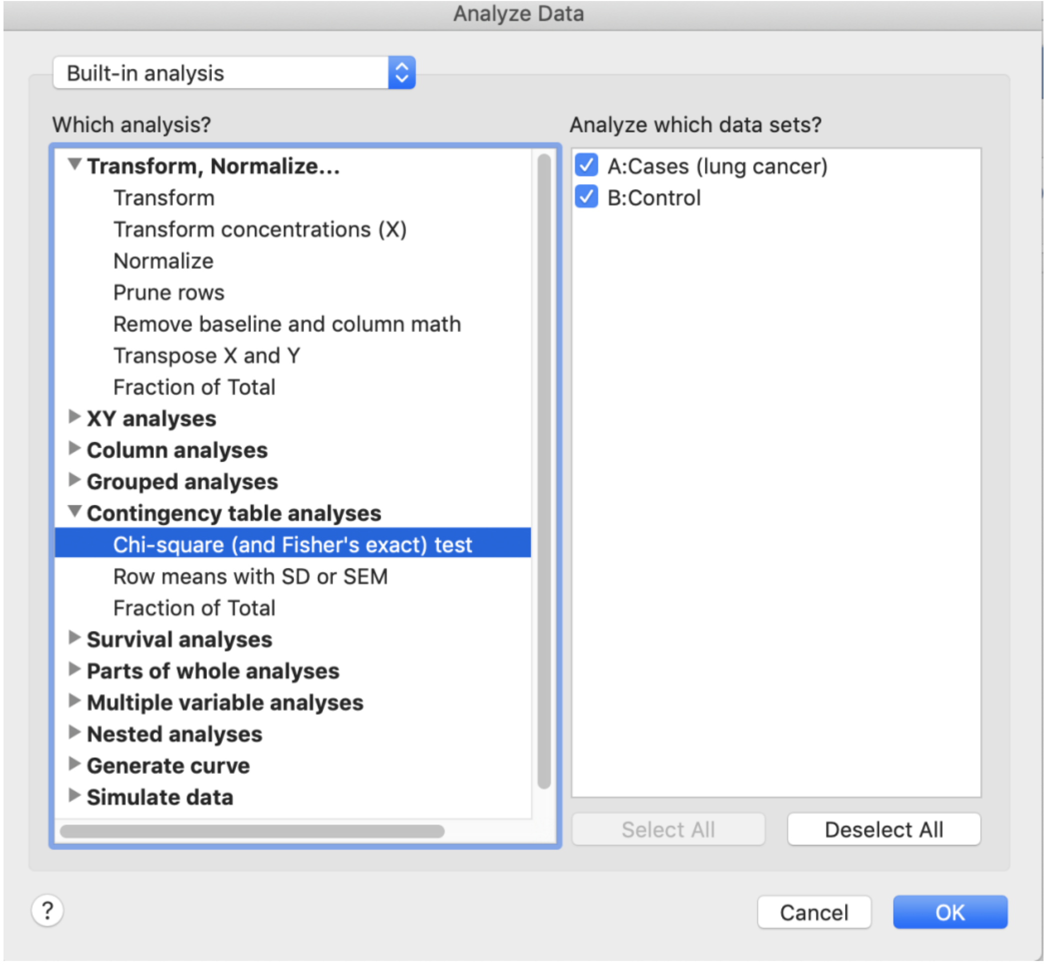


	Table format: Contingency	Outcome A	Outcome B	Outcome C
		Cases (lung cancer)	Control	Title
		Y	Y	Y
1	Smoked	688	650	
2	Never smoked	21	59	
3	Title			
4	Title			
5	Title			
6	Title			
7	Title			

3.2.3 Select **Chi-square (and Fisher's exact) test** under Contingency table analyses list. Check the columns you want to analyze on the right side, then click **OK**.



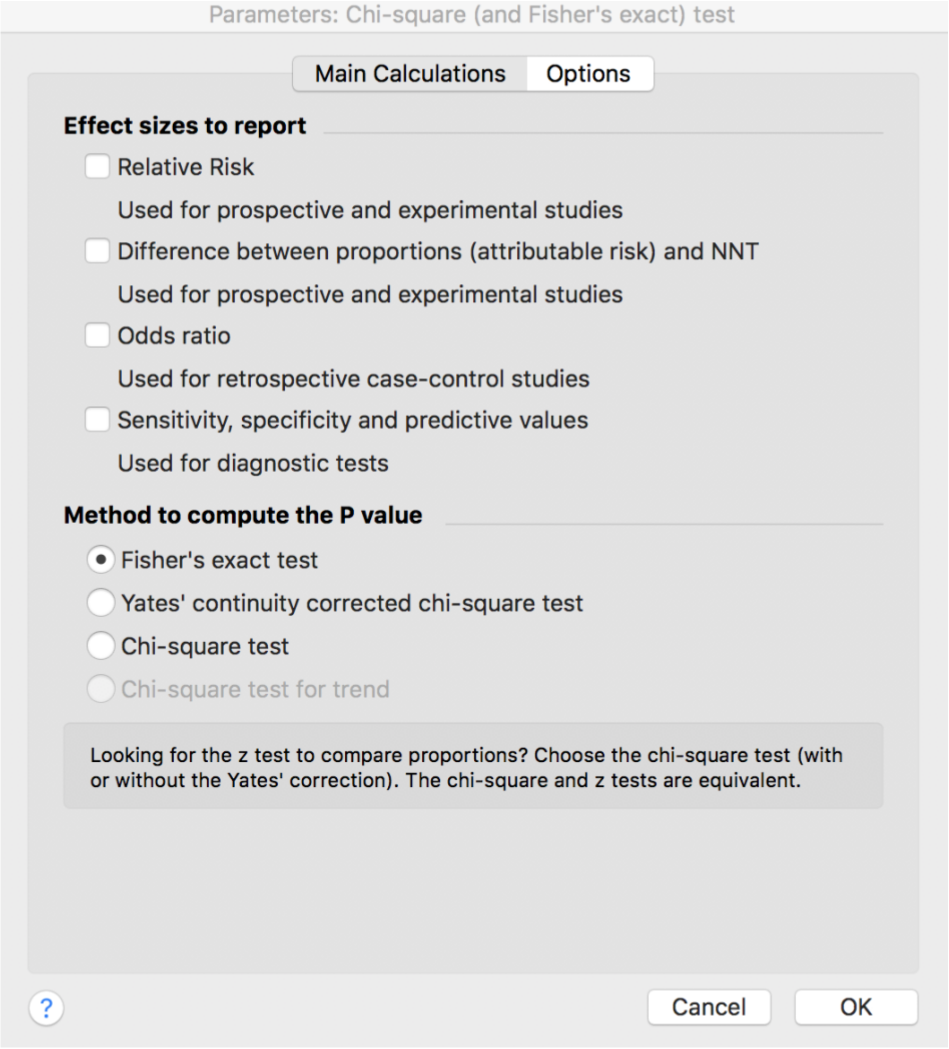
Built-in analysis

Chi-square (and Fisher's exact) test

☑ A:Cases (lung cancer)

☑ B:Control

3.2.4 Select **Fisher’s exact test**, then go to **Options**, select **Two-sided** under Calculations options, then click **OK**.



Parameters: Chi-square (and Fisher's exact) test

Effect sizes to report

□ Relative Risk

Used for prospective and experimental studies

□ Difference between proportions (attributable risk) and NNT

Used for prospective and experimental studies

□ Odds ratio

Used for retrospective case-control studies

□ Sensitivity, specificity and predictive values

Used for diagnostic tests

Method to compute the P value

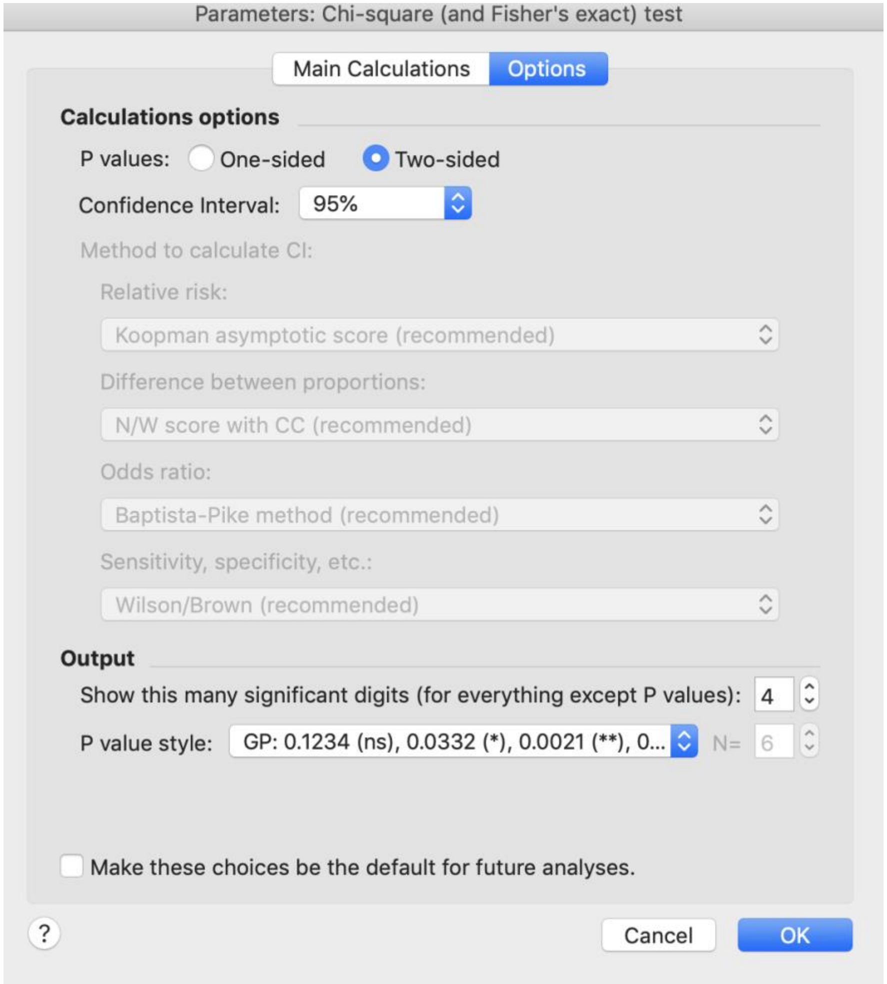
● Fisher's exact test

○ Yates' continuity corrected chi-square test

○ Chi-square test

○ Chi-square test for trend

Looking for the z test to compare proportions? Choose the chi-square test (with or without the Yates' correction). The chi-square and z tests are equivalent.



Parameters: Chi-square (and Fisher's exact) test

Calculations options

P values: ○ One-sided ● Two-sided

Confidence Interval: 95%

Method to calculate CI:

Relative risk:

Koopman asymptotic score (recommended)

Difference between proportions:

N/W score with CC (recommended)

Odds ratio:

Baptista-Pike method (recommended)

Sensitivity, specificity, etc.:

Wilson/Brown (recommended)

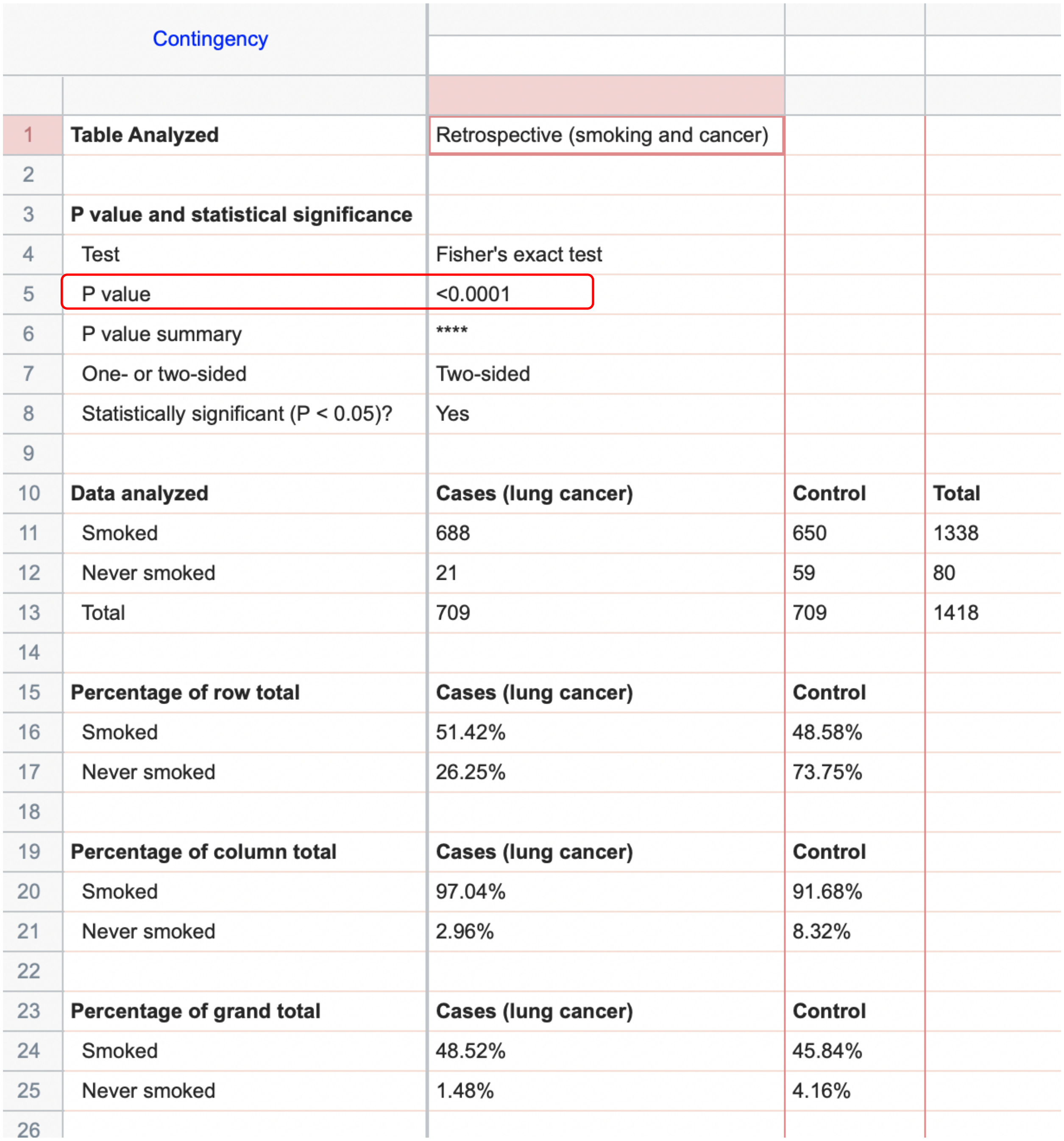
Output

Show this many significant digits (for everything except P values): 4

P value style: GP: 0.1234 (ns), 0.0332 (*), 0.0021 (**), 0... N= 6

□ Make these choices be the default for future analyses.

3.2.5 Result includes p-value, marginal table, and marginal percentage. p-value is circled. P-value is less than 0.05, we reject null hypothesis at 0.05 significance level and conclude that there is association between smoking and lung cancer.



Contingency

1	**Table Analyzed**	Retrospective (smoking and cancer)		
2				
3	**P value and statistical significance**			
4	Test	Fisher's exact test		
5	P value	<0.0001		
6	P value summary	****		
7	One- or two-sided	Two-sided		
8	Statistically significant (P < 0.05)?	Yes		
9				
10	**Data analyzed**	**Cases (lung cancer)**	**Control**	**Total**
11	Smoked	688	650	1338
12	Never smoked	21	59	80
13	Total	709	709	1418
14				
15	**Percentage of row total**	**Cases (lung cancer)**	**Control**	
16	Smoked	51.42%	48.58%	
17	Never smoked	26.25%	73.75%	
18				
19	**Percentage of column total**	**Cases (lung cancer)**	**Control**	
20	Smoked	97.04%	91.68%	
21	Never smoked	2.96%	8.32%	
22				
23	**Percentage of grand total**	**Cases (lung cancer)**	**Control**	
24	Smoked	48.52%	45.84%	
25	Never smoked	1.48%	4.16%	
26				

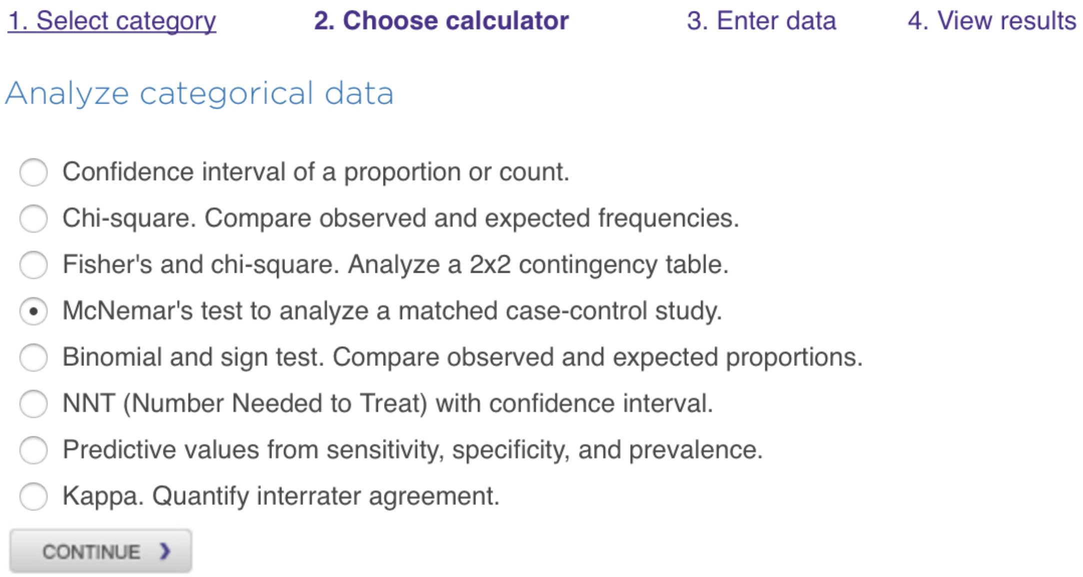
3.3 McNemar’s test

McNemar’s test is applied to paired categorical data. It is available via Prism web. (https://www.graphpad.com/quickcalcs/)

In the usual kind of case-control study, the investigator compares a group of controls with a group of cases. As a group, the controls are supposed to be similar to the cases (except for the absence of disease). Another way to perform a case-control study is to match individual cases with individual controls based on age, gender, occupation, location and other relevant variables. This is the kind of study McNemar's test is designed for.

		Control		
		+	-	Total
Case	+	13	25	38
	-	4	92	96
	Total	17	117	134

3.3.1 Click in the website above, select Categorical data, select **McNemar’s test to analyze a matched case-control study**, then click **Continue**.

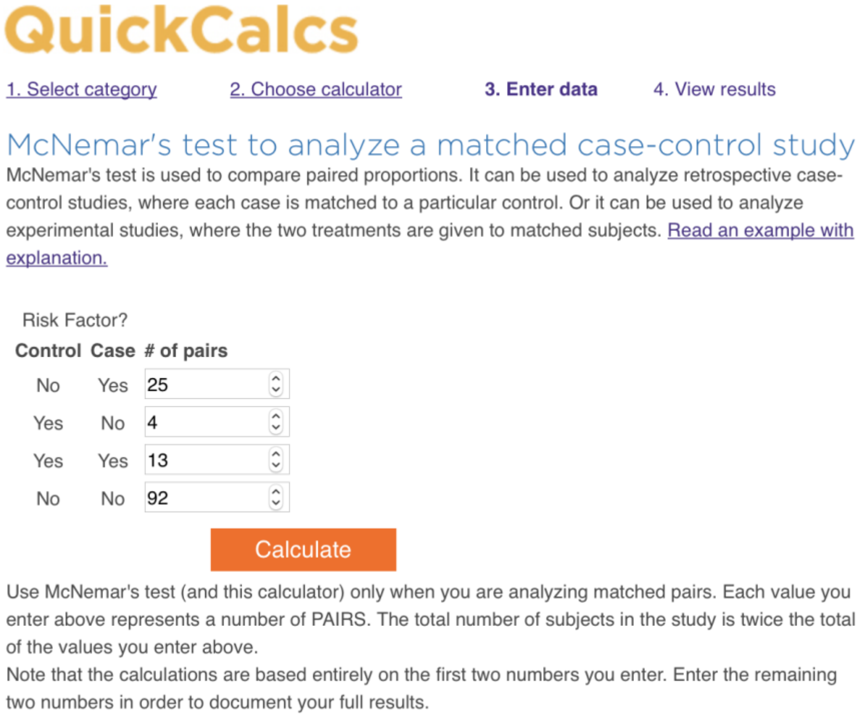


1. Select category 2. Choose calculator 3. Enter data 4. View results

Analyze categorical data

○ Confidence interval of a proportion or count.
○ Chi-square. Compare observed and expected frequencies.
○ Fisher's and chi-square. Analyze a 2x2 contingency table.
● McNemar's test to analyze a matched case-control study.
○ Binomial and sign test. Compare observed and expected proportions.
○ NNT (Number Needed to Treat) with confidence interval.
○ Predictive values from sensitivity, specificity, and prevalence.
○ Kappa. Quantify interrater agreement.

3.3.2 **Input # of pairs of case and control**. Consider "Yes" as positive, "No" as negative, input the data in the contingency table. Then click **Calculate**.



1. Select category 2. Choose calculator 3. Enter data 4. View results

McNemar's test to analyze a matched case-control study

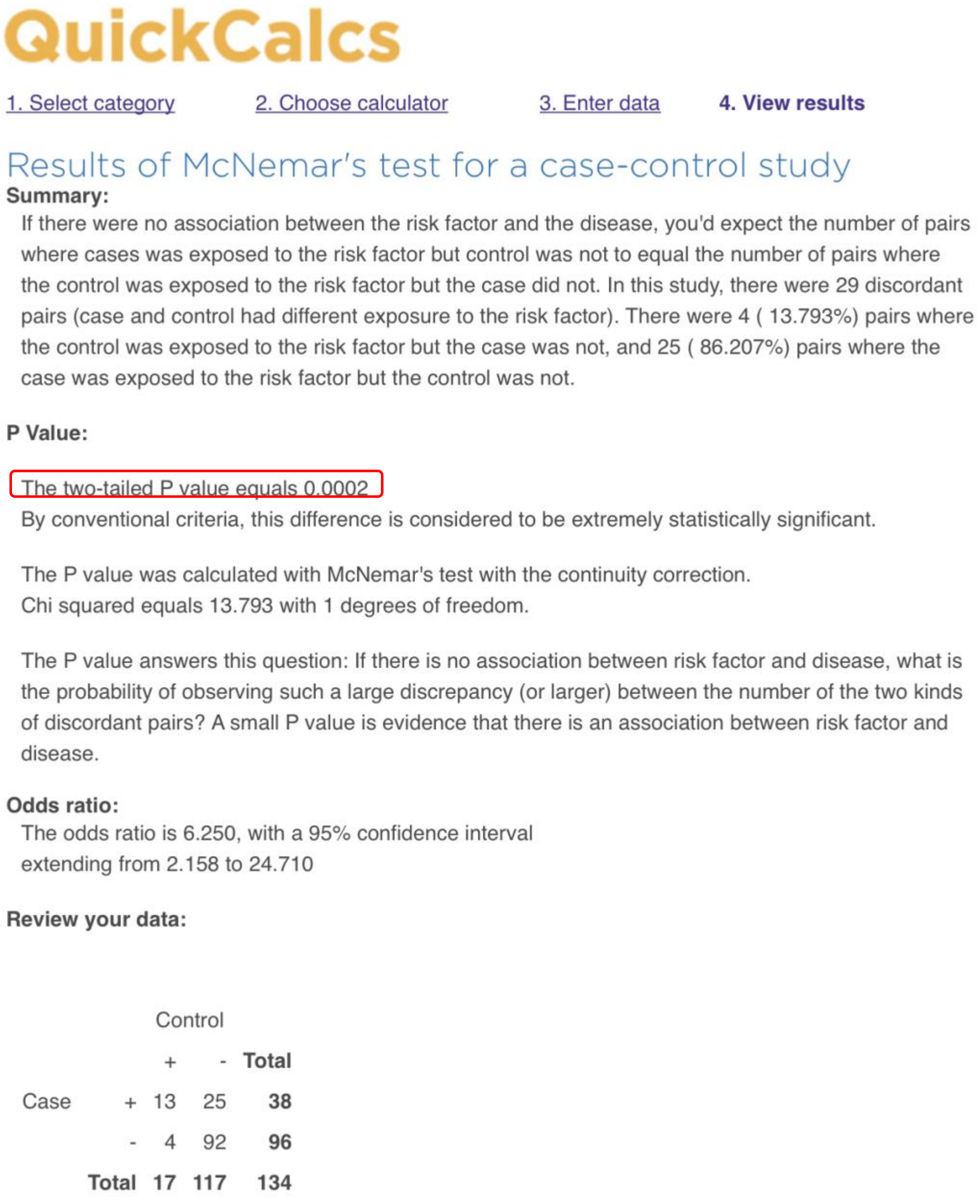
McNemar's test is used to compare paired proportions. It can be used to analyze retrospective case-control studies, where each case is matched to a particular control. Or it can be used to analyze experimental studies, where the two treatments are given to matched subjects. Read an example with explanation.

Risk Factor?		
Control	Case	# of pairs
No	Yes	25
Yes	No	4
Yes	Yes	13
No	No	92

Use McNemar's test (and this calculator) only when you are analyzing matched pairs. Each value you enter above represents a number of PAIRS. The total number of subjects in the study is twice the total of the values you enter above.

Note that the calculations are based entirely on the first two numbers you enter. Enter the remaining two numbers in order to document your full results.

3.3.3 Results include summary, p-value, odds ratio and contingency table. P-value is less than 0.05, we can reject the null hypothesis that there is association between risk factor and disease.



Results of McNemar's test for a case-control study

Summary:

If there were no association between the risk factor and the disease, you'd expect the number of pairs where cases was exposed to the risk factor but control was not to equal the number of pairs where the control was exposed to the risk factor but the case did not. In this study, there were 29 discordant pairs (case and control had different exposure to the risk factor). There were 4 (13.793%) pairs where the control was exposed to the risk factor but the case was not, and 25 (86.207%) pairs where the case was exposed to the risk factor but the control was not.

P Value:

The two-tailed P value equals 0.0002

By conventional criteria, this difference is considered to be extremely statistically significant.

The P value was calculated with McNemar's test with the continuity correction.
Chi squared equals 13.793 with 1 degrees of freedom.

The P value answers this question: If there is no association between risk factor and disease, what is the probability of observing such a large discrepancy (or larger) between the number of the two kinds of discordant pairs? A small P value is evidence that there is an association between risk factor and disease.

Odds ratio:

The odds ratio is 6.250, with a 95% confidence interval extending from 2.158 to 24.710

Review your data:

		Control		
		+	-	**Total**
Case	+	13	25	**38**
	-	4	92	**96**
	Total	**17**	**117**	**134**

In the end, if you have any questions regarding to this topic, please contact me (jingwen.gu@nih.gov) or submit a request to BCBB (bioinformatics@niaid.nih.gov).